

AI-Driven Clinical Guidance in Necrotizing Enterocolitis: Concordance with European Neonatal Care Standards

Nekrotizan Enterokolitte Yapay Zeka Destekli Klinik Rehberlik: Avrupa Yenidoğan Bakım Standartları ile Uyumun Değerlendirilmesi

© Saime SÜNDÜS UYGUN¹, © Mehmet SARIKAYA²

¹Selçuk University Faculty of Medicine, Department of Pediatrics, Division of Neonatology, Konya, Türkiye

²Selçuk University Faculty of Medicine, Department of Pediatric Surgery, Konya, Türkiye

Cite as: Sündüs Uygun S, Sarıkaya M. AI-driven clinical guidance in necrotizing enterocolitis: concordance with European Neonatal care standards. Forbes Med J. 2025;6(2):172-9

ABSTRACT

Objective: Necrotizing enterocolitis (NEC) is a life-threatening emergency in neonatal medicine, especially among premature infants, with high morbidity and mortality. Despite the availability of evidence-based resources such as the European Standards of Care for Newborn Health, variations in clinical practice persist. Artificial intelligence (AI) systems based on large language models have recently gained attention as tools to support clinical decision-making. This study aimed to evaluate the alignment of two widely used AI applications-Chat Generative Pre-trained Transformer (ChatGPT) version 4.0 and Gemini-with the European neonatal care standards in providing clinical recommendations for the management of NEC.

Methods: Forty clinical questions were prepared based on the European guidelines, covering diagnosis, treatment, nutrition, follow-up, and ethical issues. Both AI models were queried under identical conditions. Their responses were independently evaluated by a pediatric surgeon and a neonatologist using a five-point Likert scale. Inter-rater agreement and statistical comparisons were analyzed using Spearman's correlation, Cohen's kappa coefficient, and the Wilcoxon signed-rank test.

Results: ChatGPT received mean scores of 4.53 from both reviewers, while Gemini received scores of 4.33 and 4.40. Median scores for both models ranged from four to five. Spearman's correlation indicated moderate agreement between reviewers, while Cohen's kappa showed weak agreement. No statistically significant differences were found between reviewers for either model.

Conclusion: Both AI models showed high compliance with European neonatal care standards in NEC management. These findings support their potential role as supportive tools in neonatal clinical decision-making.

Keywords: Necrotizing enterocolitis, large language models, artificial intelligence, clinical guidelines, ESCNH, neonatal care, ChatGPT, Gemini

ÖZ

Amaç: Nekrotizan enterokolit (NEC), özellikle prematüre yenidoğanlarda yüksek morbidite ve mortaliteye yol açan ciddi bir klinik tablodur. European Standards of Care for Newborn Health adlı kanıta dayalı rehberle rağmen klinik uygulamalarda anlamlı düzeyde farklılıklar devam etmektedir. Son yıllarda, büyük dil modeli temelli yapay zeka (AI) uygulamaları klinik karar destek sistemleri olarak dikkat çekmektedir. Bu çalışmada, *Chat Generative Pre-trained Transformer* (ChatGPT)-4.0 ve Gemini isimli iki AI uygulamasının, NEC yönetimine yönelik olarak European Standards of Care for Newborn Health rehberi temelinde oluşturulan klinik önerilerle uyum düzeylerinin değerlendirilmesi amaçlanmıştır.

Yöntem: Söz konusu rehberle dayanarak tanı, tedavi, beslenme, izlem ve etik konularını içeren 40 açık uçlu soru hazırlanarak her iki AI uygulamasına aynı koşullarda yöneltilmiştir. Yanıtlar, bir çocuk cerrahı ve bir yenidoğan uzmanı tarafından beş puanlık Likert ölçeğiyle bağımsız şekilde değerlendirilmiştir. Değerlendiriciler arası uyum ve iki modelin karşılaştırması Spearman korelasyonu, Cohen kappa katsayısı ve Wilcoxon işaretli sıra testi ile analiz edilmiştir.

Received/Geliş: 12.05.2025

Accepted/Kabul: 04.07.2025

Yayınlanma Tarihi/

Publication Date: 06.08.2025

Corresponding Author/
Sorumlu Yazar:

Saime SÜNDÜS UYGUN, MD,

Selçuk University Faculty of
Medicine, Department of
Pediatrics, Division of Neonatology,
Konya, Türkiye

✉ uygunsaime@hotmail.com

ORCID: 0000-0002-6694-8115



Copyright© 2025 The Author. Published by Galenos Publishing House on behalf of Buca Seyfi Demirsoy Training and Research Hospital. This is an open access article under the Creative Commons AttributionNonCommercial 4.0 International (CC BY-NC 4.0) License.

Copyright© 2025 Yazar. Buca Seyfi Demirsoy Eğitim ve Araştırma Hastanesi adına Galenos Yayınevi tarafından yayımlanmıştır. Creative Commons Atıf-GayriTicari 4.0 Uluslararası (CC BY-NC 4.0) Uluslararası Lisansı ile lisanslanmış, açık erişimli bir makedir.



Bulgular: ChatGPT-4.0 her iki uzman tarafından 4,53 puan ortalamasıyla değerlendirilirken, Gemini için puanlar sırasıyla 4,33 ve 4,40 olarak bulunmuştur. Medyan puanlar her iki model için dört ile beş arasında değişmiştir. Değerlendiriciler arasında istatistiksel fark saptanmamıştır.

Sonuç: AI uygulamaları, NEC yönetiminde rehber uyumunu yüksek düzeyde sağlamış olup klinik karar süreçlerinde destekleyici araçlar olarak potansiyel taşımaktadır.

Anahtar Kelimeler: Nekrotizan enterokolit, büyük dil modelleri, yapay zeka, klinik kılavuzlar, ESCNH, yenidoğan bakımı, ChatGPT, Gemini

INTRODUCTION

Necrotizing enterocolitis (NEC) is a serious gastrointestinal emergency that primarily affects premature and low birth weight infants. The disease is characterized by intestinal inflammation, bacterial invasion, and, in advanced cases, necrosis and perforation of the intestinal wall. Despite advances in neonatal intensive care, NEC remains associated with elevated morbidity and mortality rates. In cases of NEC requiring surgical intervention, the mortality rate has been documented to exceed 30% in several series.¹ It has been shown that surviving infants are at a heightened risk of experiencing long-term complications, including but not limited to short bowel syndrome, growth retardation, and neurodevelopmental disorders.^{2,3} The unpredictable clinical course and multifactorial etiology of NEC make timely diagnosis and effective management challenging; therefore, a standards-based, evidence-based care approach is necessary.

The clinical management of NEC poses significant challenges due to the highly variable course of the disease, the potential for rapid progression, and the absence of definitive diagnostic markers. The decisions regarding medical or surgical treatment, the timing of treatment steps, and nutritional strategies can vary significantly between institutions and clinicians. Despite the existence of international guidelines, such as the European Standards for Neonatal Health (ESCNH), which provide structured frameworks for the management of NEC, there is a paucity of research on the real-time implementation of these guidelines at the bedside.⁴ This is due to limitations imposed by clinical complexities, differences in experience, and access issues. These discrepancies in implementation underscore the necessity for support systems that can seamlessly integrate clinical guideline recommendations into decision-making processes.

In recent years, there has been a marked increase in the integration of artificial intelligence (AI) into clinical applications, with its use extending to many areas, including diagnostic imaging, risk stratification, treatment planning, and clinical decision support systems.⁵ In the field of neonatal care, AI has been employed to facilitate early prediction of diseases such as sepsis, respiratory distress syndrome, and intraventricular haemorrhage in studies numbered.⁵⁻⁸ Recent developments have demonstrated the integration of AI applications into the domain of nutrition

planning, as evidenced by the implementation of a system known as TPN 2.0, which has been shown to enhance adherence to parenteral nutrition guidelines and reduce the incidence of complications in premature infants.^{9,10} Concurrently, machine learning models that amalgamate imaging and clinical data for the diagnosis of NEC have been developed and demonstrated an enhancement in diagnostic accuracy.¹¹⁻¹³ These developments demonstrate the expanding role of AI in neonatal intensive care and its potential contribution to the delivery of timely and standardised care to vulnerable patient groups.

Despite the encouraging outcomes observed in diverse domains of neonatal care, ranging from early diagnosis to nutrition planning, there is a lack of studies that systematically assess the clinical guideline alignment of LLMs in the context of NEC. Existing literature has focused more on predictive algorithms or imaging-based diagnostic systems, neglecting the assessment of textual clinical reasoning capabilities.¹¹⁻¹³ To address this need, this study aims to evaluate the degree to which large language model (LLM)-based AI applications align with established clinical guidelines in NEC scenarios. Demonstrating that such systems can produce consistent, guideline-compliant responses is critical not only for validating their safety and reliability but also for facilitating their integration into real-world neonatal care. AI systems capable of aligning with structured standards such as the ESCNH could assist clinicians by reinforcing best practices, reducing variability in care, and supporting decision-making processes in time-sensitive and resource-limited settings.

METHODS

Study Design

Both AI applications were accessed in their publicly available web-based versions: Chat Generative Pre-trained Transformer (ChatGPT)-4 (OpenAI, May 2025) and Gemini Pro 1.5 (Google Bard, accessed May 2025). No prompt engineering, temperature adjustment, or context priming was used. Each of the 40 clinical questions was presented as a plain-text prompt during a single uninterrupted session per model. The ESCNH guideline document was not uploaded or attached to the prompt; the AI models were expected to generate responses based on their internal training data and knowledge up to the date of access. To ensure standardization, each question was

asked only once per model, without regeneration or multiple attempts. All responses were recorded in their original form, reflecting a single-use, real-time interaction to simulate a realistic clinical consultation scenario. No post-processing or modification of AI responses was performed. To minimize evaluator bias, the responses from ChatGPT and Gemini were anonymized and randomly ordered before being presented to the reviewers. Both evaluators were blinded to the source of the AI responses during the scoring process. To enhance transparency and clinical relevance, one example case-including the prompt and responses generated by both AI systems-has been provided in the supplementary material (Appendix A).

A total of 40 open-ended clinical questions were created based on the official recommendations of the European Standards of Care for ESCNH, covering core areas such as diagnosis, treatment, follow-up, communication, nutrition, and ethical considerations in NEC care. These questions were presented separately to both AI applications under identical conditions, and each response was recorded in its original form without modification.

The responses generated by each AI application were then independently assessed by two specialists: a paediatric surgeon and a neonatologist, both with more than 10 years of clinical experience in neonatal care. The reviewers scored each AI response using a structured 5-point Likert scale to determine its level of agreement with the ESCNH guideline recommendations. The 5-point Likert scale used by the reviewers was structured as follows: 1 = not compliant with the guideline, 2 = low compliance, 3 = moderate compliance, 4 = high compliance, and 5 = fully compliant with the guideline. This methodological design was chosen to simulate real-world use scenarios and to assess the comparative performance of AI models in providing evidence-based clinical guidance.

Statistical Analysis

All statistical analyses were conducted using IBM SPSS Statistics version 29.0 (IBM Corp., Armonk, NY). Descriptive statistics were presented as medians with interquartile ranges (IQR), in accordance with the non-parametric nature of the data. Inter-rater agreement was assessed using Spearman's rank correlation coefficient (ρ) for ordinal data and Cohen's kappa (κ) for categorical agreement. The Wilcoxon signed-rank test was used to compare the Likert scores assigned by the pediatric surgeon and the neonatologist for each AI system, as the data were not normally distributed. A p value <0.05 was considered statistically significant.

Ethical Consideration

Since this study did not involve human or animal subjects and was limited to AI-based textual response evaluation, it was exempt from ethics committee approval. As the evaluated responses were AI-generated and did not involve patient data or human participation, the study posed no ethical risk.

RESULTS

The mean Likert score for ChatGPT responses was 4.53 ± 0.72 based on pediatric surgeon evaluations and 4.53 ± 0.64 based on neonatologist ratings. Similarly, Gemini responses received an average score of 4.33 ± 0.69 from the pediatric surgeon and 4.40 ± 0.55 from the neonatologist. Median values for both AI systems were 5, with an IQR of 4-5 for both reviewers across all models.

A moderate and statistically significant correlation was observed between the two reviewers' Likert scores for both AI systems (Figure 1 for Gemini, Figure 2 for ChatGPT). For ChatGPT, Spearman's ρ was 0.41 ($p=0.008$), while for Gemini, the correlation coefficient was 0.35 ($p=0.027$). These findings suggest reasonable consistency in expert assessments, despite minor variations.

Cohen's kappa statistic revealed weak categorical agreement between the reviewers, with $\kappa=0.10$ for ChatGPT and $\kappa=0.21$ for Gemini. This low kappa is likely due to the ordinal nature of the Likert scale rather than true disagreement.

Although the vast majority of responses received scores of 4 or 5, a few instances of moderate compliance (score=3) were observed. These cases often involved complex, context-sensitive issues such as ethics or multidisciplinary coordination, where even human experts may interpret guideline recommendations differently. Such variability underscores the importance of expert oversight when integrating AI tools into clinical workflows.

Wilcoxon signed-rank tests demonstrated no statistically significant differences in Likert scores between the pediatric surgeon and neonatologist for either AI system (Table 1). None of the AI responses received a score of 1 (not compliant) or 2 (low compliance) from either expert. Only 3 of the total 80 ratings (3.75%) were scored as 3 (moderate compliance), reflecting some ambiguity in guideline interpretation in specific ethical or interdisciplinary domains. The p value was 1.000 for ChatGPT and 0.513 for Gemini; however, this does not necessarily indicate strong consistency in reviewer scoring within each AI system. Although ChatGPT consistently received slightly higher scores than Gemini, the differences were not statistically significant for either reviewer ($p=0.133$ for

pediatric surgeon; $p=0.219$ for neonatologist), indicating comparable levels of guideline adherence between the two AI systems (Table 2).

DISCUSSION

In this prospective comparative study, two widely used LLMs, ChatGPT and Gemini, were evaluated for their adherence to the ESCNH in the context of NEC.

Both systems demonstrated a high level of guideline compliance, with ChatGPT achieving slightly higher mean Likert scores than the two expert raters. Notably, 96.25% of the expert scores fell within the 4-5 range on the Likert scale, reflecting a consistently high or full compliance with ESCNH recommendations across both AI platforms. These findings highlight the potential of LLMs as decision support tools in neonatal care, particularly in the management of

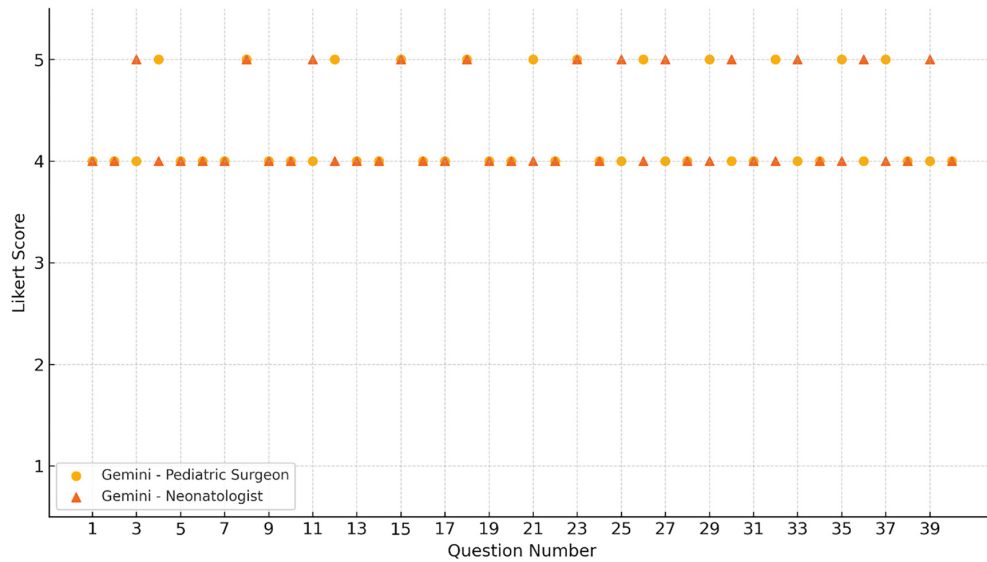


Figure 1. Reviewer scores for Gemini across 40 clinical questions. Scatterplot shows Likert scores assigned by the pediatric surgeon and neonatologist for each Gemini-generated response. Scores range from 1 (not compliant) to 5 (fully compliant) based on ESCNH guideline adherence

ESCNH: European Standards for Neonatal Health

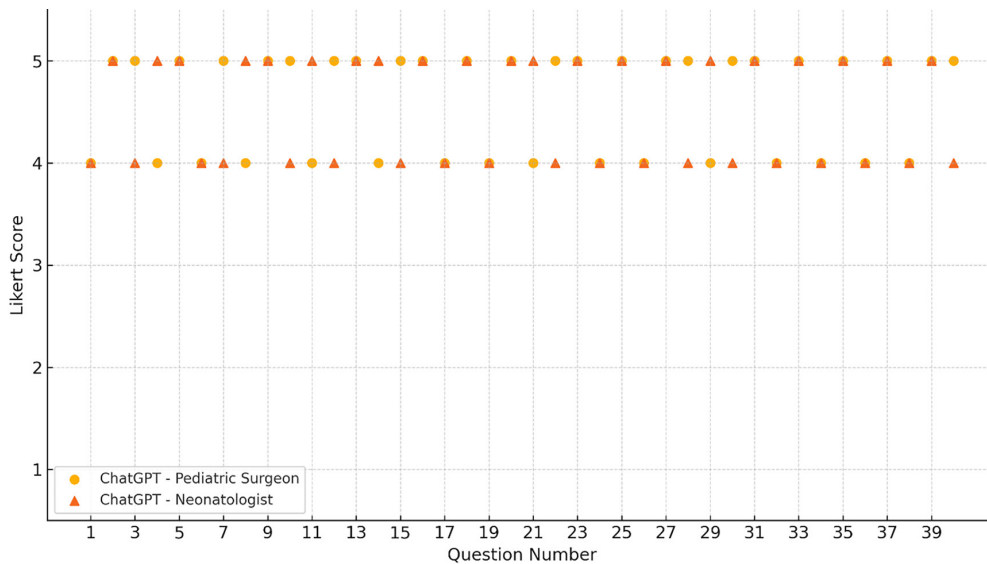


Figure 2. Reviewer scores for ChatGPT across 40 clinical questions. Scatterplot illustrates the distribution of Likert scores assigned by the pediatric surgeon and neonatologist for ChatGPT-generated responses. Responses were evaluated for alignment with ESCNH recommendations

ChatGPT: Chat Generative Pre-trained Transformer, ESCNH: European Standards for Neonatal Health

Table 1. Reviewer scores and inter-rater agreement metrics for ChatGPT and Gemini					
AI system	Reviewer	Median (IQR)	Spearman ρ (p)	Cohen's κ	Wilcoxon p
ChatGPT	Pediatric surgeon	5 (4-5)	0.41 (p=0.008)	0.10	1.000
ChatGPT	Neonatologist	5 (4-5)			
Gemini	Pediatric surgeon	4 (4-5)	0.35 (p=0.027)	0.21	0.513
Gemini	Neonatologist	4 (4-5)			
(Scoring based on a 5-point Likert scale: 1 = not compliant, 2 = low compliance, 3 = moderate, 4 = high, 5 = fully compliant with ESCNH guidelines) ChatGPT: Chat Generative Pre-trained Transformer, IQR: Interquartile range					

Table 2. Comparison of mean scores for ChatGPT and Gemini by Reviewer				
Reviewer value	ChatGPT median (IQR)	Gemini median (IQR)	Mann-Whitney U	p value
Pediatric surgeon	5 (4-5)	4 (4-5)	940.00	0.133
Neonatologist	5 (4-5)	4 (4-5)	913.50	0.219
(Higher scores reflect stronger alignment with ESCNH guidelines) ChatGPT: Chat Generative Pre-trained Transformer, IQR: Interquartile range, ESCNH: European Standards for Neonatal Health				

complex and time-sensitive situations where adherence to standard protocols, such as NEC, is critical.

Furthermore, the moderate inter-rater agreement observed using Spearman’s correlation suggests a reassuring level of consistency in expert judgement. Although Cohen’s kappa values were relatively low, this is likely to be due to methodological limitations associated with the application of categorical agreement statistics to ordinal data, rather than a genuine lack of consensus. Overall, the findings position LLMs as promising adjunct tools in neonatal clinical reasoning when linked to structured clinical frameworks such as ESCNH guidelines.

While numerous studies have explored applications of AI in general medicine and adult subspecialties, relatively few studies have systematically evaluated LLMs in neonatal care using structured, guideline-based assessments. For example, Singer et al.⁹ and Phongpreecha et al.¹⁰ reported encouraging results with AI systems in neonatal nutrition planning and precision parenteral nutrition strategies, but these studies focused on algorithm-driven interventions rather than textual clinical reasoning. Similarly, Sullivan et al.⁷ highlighted the role of AI in predicting neonatal sepsis and respiratory failure through data modeling, but these approaches did not include direct guideline adherence.⁶

In contrast, our study offers several novel directions: it addresses a critical NEC, it uses a rigorously developed question set fully derived from the ESCNH guidelines, and it uses a dual expert assessment model involving both a paediatric surgeon and a neonatologist. To our knowledge, this is the first study to evaluate guideline concordance of LLM-generated clinical responses specifically in the context of NEC.

Previous research on AI-assisted NEC diagnosis has primarily focused on image-based tools, such as deep learning models for interpreting abdominal radiographs or integrating clinical data with imaging for early diagnosis.¹¹⁻¹³ While these approaches are valuable, they do not capture the narrative reasoning and context-sensitive decision making that is central to LLM performance. Therefore, our findings help to fill a critical gap by assessing how well LLMs reflect structured neonatal guideline recommendations through textual outputs. In addition, Sarikaya et al.¹⁴ evaluated the guideline compatibility of multiple AI platforms in the management of vesicoureteral reflux and found that LLMs showed high levels of agreement with pediatric urology guidelines, strengthening their applicability in structured pediatric decision-making frameworks.

The structured nature of this study enhances both the content validity and the direct applicability of its findings to neonatal clinical decision-making. By formulating all questions strictly based on ESCNH recommendations, the study reflects practical scenarios that clinicians frequently encounter in neonatal intensive care settings. In addition, presenting the same set of questions to both AI models under identical conditions eliminates variability, and the blinded assessment of anonymized responses by two independent experts helps minimize contextual and cognitive biases. This methodology increases the internal validity of the results and supports the comparability of AI performance across platforms—an essential factor when considering their future integration into standardized care protocols or decision-support systems in NICUs. Second, the inclusion of two independent raters (a pediatric surgeon and a neonatologist) from different but complementary specialties provides a multidisciplinary perspective and increases the generalizability of the results across neonatal

care settings. Third, the prospective and standardized assessment design, including same-day referral and the same set of questions for both LLMs, minimizes potential contextual bias and ensures comparability.

However, certain limitations should also be recognized. Although widely used in health informatics research, the 5-point Likert scale used to assess guideline adherence is inherently subject to subjective interpretation, which may affect inter-rater reliability measures such as Cohen's kappa. Although Spearman's correlation indicates a moderate level of agreement between raters, low kappa values may reflect the ordinal nature of the scale rather than significant disagreement. In addition, only two LLM-based systems were evaluated, limiting generalizability to other existing or new models.

The results of this study support the potential integration of LLM-based systems into neonatal clinical settings, particularly in roles involving decision support, guidelines training and clinical education. Given the high level of alignment with ESCNH recommendations demonstrated by both ChatGPT and Gemini, such tools may be particularly useful in settings where access to neonatal subspecialists is limited or where rapid dissemination of standardized information is required. LLMs can be incorporated into electronic health records as real-time, just-in-time decision support tools that reinforce evidence-based practice and reduce variability in care.

However, caution should be exercised when interpreting these findings for real-world application. Although the LLMs evaluated provided largely guideline-compliant responses, occasional inconsistencies or omissions were noted, highlighting the ongoing need for expert review and validation. Future research should aim to develop domain-specific LLMs trained on neonatal data, validate them in different clinical settings, and explore their impact on diagnostic accuracy, clinician efficiency, and parental satisfaction. In addition, ethical considerations, including transparency, privacy, and liability issues in AI-assisted clinical decision-making, should be systematically addressed before widespread implementation.

This study provides the first evidence that LLMs can comply with neonatal clinical guidelines in complex scenarios such as necrotising enterocolitis, providing a promising basis for future AI-assisted care in neonatology.

The integration of LLM-based systems into NICU workflows may offer several practical advantages, such as providing real-time clinical decision support, reinforcing adherence to standardized guidelines, and serving as educational tools for junior clinicians or training simulations. Particularly in resource-limited or high-pressure settings, AI systems may act as supplementary aids to improve consistency in

care. However, their implementation must be approached cautiously. Medicolegal concerns—such as liability in case of AI-related error, the interpretability of outputs, and the transparency of data sources—must be addressed through clear institutional policies. Furthermore, ethical considerations, including patient safety, data privacy, and clinician responsibility, remain critical to responsible deployment. Continuous human oversight should be maintained to validate AI outputs and ensure alignment with patient-specific contexts.

The low Cohen's kappa values observed in our study may suggest weak inter-rater agreement. However, this likely stems from the statistical limitations of using kappa with ordinal data, especially when most ratings are concentrated at the high end of the scale (i.e., scores of 4 and 5). In such skewed distributions, even minor differences can disproportionately affect the kappa coefficient. To provide a more appropriate measure, we also reported Spearman's rank correlation, which demonstrated moderate and statistically significant agreement between reviewers. Although other methods such as Gwet's ACI or intraclass correlation coefficient could be employed, the 5-point Likert scale remain widely used in clinical guideline adherence studies and was selected for its clarity, ease of interpretation, and suitability for the scoring task assigned to expert raters.

Study Limitations

This study's findings should be interpreted with caution due to several limitations. Primarily, the evaluation was confined to two LLM-based AI systems, ChatGPT-4 and Gemini, which may restrict the generalizability of the results to other existing or forthcoming models. Secondly, each clinical prompt was presented to the AI systems only once, without employing multiple iterations or prompt refinement strategies. This approach may have influenced the observed response variability. Lastly, the ESCNH guideline was not directly integrated or linked to the AI systems. Consequently, the assessment reflects the models' inherent knowledge bases as of May 2025, which are subject to continuous evolution. Future research endeavors should aim to incorporate a broader spectrum of LLMs, simulate real-time clinical scenarios, and include a more comprehensive evaluation of patient outcomes to further validate these findings.

CONCLUSION

This study provides preliminary evidence that LLMs, such as ChatGPT-4 and Gemini, can generate clinical responses that align closely with the ESCNH in the context of NEC. The findings demonstrate high levels of guideline compliance, particularly in structured clinical scenarios,

and suggest that LLMs hold potential as supplementary decision-support tools in neonatal intensive care units.

While limitations exist—including the use of a single response per prompt, the inherent variability of AI outputs, and the reliance on Likert-based expert scoring—these do not compromise the study's overall validity. Instead, they highlight the importance of continued validation, transparent implementation, and responsible oversight.

As AI models continue to evolve, future studies should explore their integration into clinical workflows, training environments, and decision-making frameworks, ensuring that such technologies enhance, rather than replace, expert judgment in neonatal care.

Ethics

Ethics Committee Approval and Informed Consent:

This study was exempt from ethics committee approval because it did not involve human or animal subjects and was limited to AI-based textual response evaluation. No ethics committee approval or patient consent was required for the study, as the evaluated responses were generated by AI and did not involve patient data or human participation.

Footnotes

Authorship Contributions

Concept: S.S.U., Design: M.S., Data Collection or Processing: M.S., Analysis or Interpretation: S.S.U., Literature Search: M.S., Writing: S.S.U.

Conflict of Interest: No conflict of interest was declared by the authors.

Financial Disclosure: The authors declared that this study received no financial support.

REFERENCES

1. Fitzgibbons SC, Ching Y, Yu D, et al. Mortality of necrotizing enterocolitis expressed by birth weight categories. *J Pediatr Surg*. 2009;44:1072-5.

2. Patel RM, Kandefer S, Walsh MC, et al. Causes and timing of death in extremely premature infants from 2000 through 2011. *N Engl J Med*. 2015;372:331-40.
3. Neu J, Walker WA. Necrotizing enterocolitis. *N Engl J Med*. 2011;364:255-64.
4. Lindacher V, Altebaeumer P, Marlow N, et al. European Standards of Care for Newborn Health—a project protocol. *Acta Paediatr*. 2021;110:1433-8.
5. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med*. 2019;25:44-56.
6. Sullivan BA, Kausch SL, Fairchild KD. Artificial and human intelligence for early identification of neonatal sepsis. *Pediatric Res*. 2023;93:350-6.
7. Sullivan BA, Beam K, Vesoulis ZA, et al. Transforming neonatal care with artificial intelligence: challenges, ethical consideration, and opportunities. *J Perinatol*. 2024;44:1-11.
8. De Rose DU, Ronchetti MP, Martini L, et al. Diagnosis and management of neonatal bacterial sepsis: current challenges and future perspectives. *Trop Med Infect Dis*. 2024;9:199.
9. Singer P, Robinson E, Raphaelli O. The future of artificial intelligence in clinical nutrition. Current opinion in clinical nutrition and metabolic care. *Curr Opin Clin Nutr Metab Care*. 2024;27(2):200-6.
10. Phongpreecha T, Ghanem M, Reiss JD, et al. AI-guided precision parenteral nutrition for neonatal intensive care units. *Nat Med*. 2025;31:1882-94.
11. Sitek A, Seliga-Siwecka J, Plotka S, et al. Artificial intelligence in the diagnosis of necrotising enterocolitis in newborns. *Pediatric Res*. 2023;93:376-81.
12. Duci M, Verlato G, Moschino L, Uccheddu F, Fascetti-Leon F. Advances in artificial intelligence and machine learning for precision medicine in necrotizing enterocolitis and neonatal sepsis: a state-of-the-art review. *Children (Basel)*. 2025;12:498.
13. Weller JH, Scheese D, Tragesser C, Yi PH, Alaish SM, Hackam DJ. Artificial intelligence vs. doctors: diagnosing necrotizing enterocolitis on abdominal radiographs. *J Pediatr Surg*. 2024;59:161592.
14. Sarikaya M, Ozcan Siki F, Ciftci I. Use of artificial intelligence in vesicoureteral reflux disease: a comparative study of guideline compliance. *J Clin Med*. 2025;14:2378

Appendix A.

Sample Case Scenario and AI Responses

Prompt

"A 28-week premature infant with abdominal distension, bile-stained gastric residuals, and signs of systemic instability is suspected of having necrotizing enterocolitis (NEC). According to European Standards for Neonatal Health (ESCNH) guidelines, how should this patient be managed in terms of diagnostics, nutrition, and surgical consultation?"

ChatGPT-4b Response

Based on the ESCNH guidelines, the management of a 28-week premature infant suspected of having NEC should include:

Diagnostics

- Immediate cessation of enteral feeds.
- Serial abdominal examinations and measurements.
- Abdominal X-ray to identify pneumatosis intestinalis, portal venous gas, or pneumoperitoneum.
- Laboratory testing including complete blood count, C-reactive protein, blood culture, and serum electrolytes.

Nutrition

- Initiation of total parenteral nutrition (TPN).
- Monitoring of fluid balance and metabolic parameters.

Surgical Consultation

- Early involvement of the pediatric surgery team.
- Indications for surgical evaluation include abdominal wall discoloration, fixed bowel loop on imaging, and clinical deterioration despite conservative measures.

Close multidisciplinary collaboration is essential, and transfer to a tertiary neonatal surgical center should be considered if surgical capacity is limited locally.

Gemini Response

For a 28-week gestational age infant showing signs suggestive of NEC, ESCNH guidelines recommend the following approach:

Diagnostic Approach

- Discontinue oral or tube feeding.
- Perform serial abdominal radiographs to detect early signs such as pneumatosis intestinalis.
- Laboratory evaluation with blood counts, inflammatory markers, and metabolic panels.

Nutritional Support

- Initiate TPN and maintain adequate hydration and electrolyte balance.
- Withhold enteral feeds until clinical and radiologic improvement is observed.

Surgical Considerations

- Consult pediatric surgery early.
- Signs like abdominal rigidity, erythema, metabolic acidosis, and free air are key indicators for potential surgical intervention.

A team-based approach involving neonatologists, radiologists, and pediatric surgeons is emphasized to ensure timely decision-making.