










ChatGPT for clinical use in labor management: A prospective cohort study

 ¹Ali Selçuk YENİOCAK
 ¹Can TERCAN
 ¹Emrah DAĞDEVİREN
 ¹Emrullah AKAY
 ¹Deniz ARAS
 ¹Seda MAŞ
 ²Eralp BULUTLAR
 ³Gizem Berfin ULUUTKU BULUTLAR
 ⁴Süleyman SALMAN

¹Department of Obstetrics and Gynecology, Başakşehir Çam and Sakura City Hospital, Istanbul, Turkey

²Department of Obstetrics and Gynecology, University of Health Sciences, Turkey. Istanbul Zeynep Kamil Maternity and Children's Diseases Health Training and Research Center, Istanbul, Turkey

³Department of Obstetrics and Gynecology, Haydarpaşa Numune Training and Research Hospital, Istanbul, Turkey

⁴Department of Obstetrics and Gynecology, Gaziosmanpaşa Taksim Training and Research Hospital, Istanbul, Turkey

ORCID ID

ASY : 0000-0002-8149-6348
CT : 0000-0003-1325-6294
ED : 0000-0002-1730-3724
EA : 0000-0003-3792-7777
DA : 0000-0001-6018-3206
SM : 0009-0007-3228-0413
EB : 0000-0002-2246-4899
GBUB : 0000-0001-6979-0854
SS : 0000-0001-7090-6105



ABSTRACT

Objective: Artificial intelligence, particularly machine learning, has shown promise in medical applications. This study evaluates the diagnostic accuracy and generalizability of the large language model ChatGPT4.0 in predicting labor protraction.

Material and Methods: A prospective, single-center cohort study analyzed retrospective data from 100 term pregnancies at low risk for labor protraction. The sample size was calculated using G*Power for 95% statistical power (minimum 46 patients). ChatGPT4.0 was tested on identifying 14 cesarean cases due to labor protraction and predicting active labor durations. The process was repeated after one week to assess consistency. Statistical analyses included Kolmogorov-Smirnov, Mann-Whitney U, Fisher's Exact, Friedman's, and independent t-tests ($p < 0.05$ significance).

Results: ChatGPT4.0 achieved 80% overall diagnostic accuracy, with 28.57% sensitivity and 88.37% specificity at initial and follow-up predictions ($p = 0.105$). However, predicted labor durations significantly differed from real-world data: initial (3.66 ± 1.69 hours), follow-up (6.23 ± 0.50 hours), and actual (5.17 ± 2.80 hours) ($p < 0.001$). The difference between initial and follow-up predictions was statistically insignificant ($p = 0.388$).

Conclusion: ChatGPT4.0 demonstrates high specificity in identifying labor protraction risks but shows inconsistencies in prediction accuracy, raising concerns about reliability and generalizability. Further research is needed to refine AI tools for clinical applications while ensuring ethical and safety standards. AI has potential in obstetric decision-making but requires rigorous evaluation before integration into practice. The significant limitation of ChatGPT is its restricted generalizability, largely due to the "black box" nature of the algorithm.

Keywords: AI, artificial intelligence, ChatGPT4.0, labor, large language model, LLM, machine learning, ML, obstetrics, protraction.

Cite this article as: Yeniocak AS, Tercan C, Dağdeviren E, Akay E, Aras D, Maş S, et al. ChatGPT for clinical use in labor management: A prospective cohort study. Zeynep Kamil Med J 2025;56(3):119–126.

Received: February 04, 2025 **Revised:** May 13, 2025 **Accepted:** May 20, 2025 **Online:** August 15, 2025

Correspondence: Ali Selçuk YENİOCAK, MD. Başakşehir Çam ve Sakura Şehir Hastanesi, Kadın Hastalıkları ve Doğum Kliniği, İstanbul, Türkiye.

Tel: +90 532 510 11 80 **e-mail:** a.s.yeniocak@hotmail.com

Zeynep Kamil Medical Journal published by Kare Publishing. Zeynep Kamil Tıp Dergisi, Kare Yayıncılık tarafından basılmıştır.

OPEN ACCESS This is an open access article under the CC BY-NC license (<http://creativecommons.org/licenses/by-nc/4.0/>).



INTRODUCTION

Artificial intelligence (AI) is defined in the Oxford Dictionary as “the theory and development of computer systems able to perform tasks that normally require human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages.”^[1] The question of whether it is logically possible to say that a machine can think was first addressed in 1950 in the philosophy magazine *Mind*. This was in the famous article *Computing Machinery and Intelligence* by the British mathematician and computer scientist Alan Turing, who proposed the Turing Test.^[2] Six years later, John McCarthy used the term AI for the previously named “thinking machines” and proposed AI as a new scientific field.^[3] Despite all the scientific interest and developments, the acceptance and widespread use of AI took place in the early 2000s with the establishment of machine learning (ML) algorithms.

In recent years, AI has become a part of our daily routines, surrounding nearly every aspect in both visible and invisible ways, especially with the application of artificial neural networks (ANNs).^[4] An artificial neural network can be defined as a massively parallel combination of simple processing units that acquire knowledge from the environment through a learning process and store the knowledge and its connections.^[5] By mimicking a human nervous system—comprising the neuron’s body, axon, dendrites, and synapses—ANN processing elements capture information as synaptic weight. They combine input signals using a summing function and calculate the output through an activation function. The ML of ANNs is defined as interconnecting with other processing elements and modifying synaptic weight to capture new information, resulting in changes to its own topography.^[6]

The application of AI technologies has become increasingly popular in medical research in recent years; however, the daily use of AI is not as new as it seems, especially in obstetrics and gynecology. Building 3D images from 2D planned ultrasonography images has almost become a part of daily obstetric practice.^[7] Recent studies on breast screening and early diagnosis of breast cancer using ANNs have shown promising results.^[8] In addition, studies on the use of ANNs in the diagnosis and classification of adnexal masses are important scientific developments that may help clinicians make objective diagnoses in the future.^[9] Moreover, not only imaging methods but also many promising studies—from selecting the most viable embryo in *in vitro* fertilization to endometriosis or preeclampsia diagnostic algorithms, and the use of AI in robotic surgery—have been published recently.^[10,11]

OpenAI launched a large language model (LLM) AI, ChatGPT (Chat Generative Pre-Trained Transformer), on November 20, 2022.^[12] However, AI-generated output accuracy and other legal and ethical concerns have not been fully addressed yet. Despite this, ChatGPT reached more than 100 million users within the first few months after its launch^[12] and is seen by many as the future of medicine.

Objectives

The primary objective of our study is to investigate the diagnostic accuracy of ChatGPT in predicting labor protraction by assessing the risk of intrapartum cesarean section within the study cohort. The secondary objective is to investigate the generalizability of the results generated by AI by comparing initial and one-week

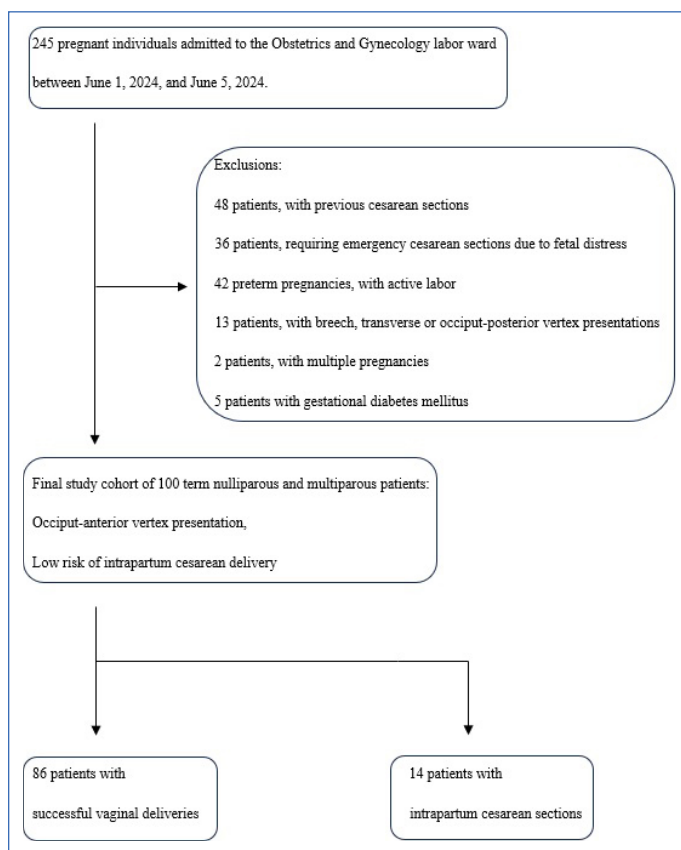


Figure 1: Time chart of the study.

follow-up responses generated by ChatGPT 4.0 for the same questions and study cohort.

MATERIAL AND METHODS

Ethics Approval and Consent to Participate

The study was conducted after receiving approval from the Basaksehir Cam and Sakura City Hospital Clinical Investigations Ethics Committee on December 10, 2024 (Ethics number: KAEK-11/30.10.2024.222), in accordance with the principles of the Declaration of Helsinki. This study was registered at ClinicalTrials.gov (Identifier: NCT KAEK-11/30.10.2024.222). Due to the retrospective nature of the study, no additional consent was obtained from the participants. However, all patients provided written informed consent upon admission for the use of their clinical records in scientific research, with a guarantee of anonymity, as approved by the local ethics committee. Consent to Participate declaration: not applicable. All patients had written informed consent upon their admission stating that their clinical records might be utilized for scientific research with a guarantee of anonymity, approved by the local ethics committee.

Study Design

This prospective single-center cohort study aimed to investigate the intrapartum risk assessment of our retrospective study cohort by AI (ChatGPT4.0). A detailed time chart of the study population is shown in Figure 1.

ChatGPT, I'll paste here a data of 100 patients. None of them are multiple pregnancies, gestational diabetes, non-vertex presentations or previous cesarean sections. all of them have a cervical dilatation of 6cm with effective uterine contractions (active phase of labor). We follow ACOG's recommendations for managing the second stage of labor, which includes individualized assessment to tailor care to each patient. To optimize labor outcomes, amniotomy is performed during augmentation or induction to reduce the duration of labor. Both low-dose and high-dose oxytocin protocols are considered effective for active labor management, helping to minimize the need for operative deliveries. . Active Phase Protraction and Arrest Disorder is diagnosed in patients with ruptured membranes and at least 6 cm dilation if there is no cervical progression after 4 hours of adequate uterine activity. Prolonged second stage of labor is diagnosed when pushing exceeds 3 hours in nulliparous individuals or 2 hours in multiparous individuals, although diagnosis is based on individual circumstances. In cases of active phase arrest, cesarean delivery is performed. 86 of the patients had successful vaginal delivery, 14 of them had a cesarean section due to 'Active Phase Protraction and Arrest Disorder'. I want you to triage most risky 14 patients who may have cesarean sections? and I want you to guess the active labor time (in hours) for the rest of the 86.

Figure 2: ChatGPT4.0 was asked to identify the 14 patients at highest risk for intrapartum cesarean section.

Demographic data including patient age, gestational age (in weeks), gravidity and parity numbers, ultrasound evaluation of estimated fetal weight (EFW) upon labor ward admission, patients' body mass index (BMI), and active labor time were collected and analyzed. A table of the study cohort's demographic data with blinded patient names, identification numbers, active labor times, and intrapartum cesarean section cases was uploaded to ChatGPT4.0 (Appendix 1). To investigate the diagnostic accuracy of the AI, ChatGPT4.0 was asked to identify the 14 patients at highest risk for labor protraction (Fig. 2).

After this step, ChatGPT4.0 was asked to predict active labor times in hours for each patient in the study cohort. To investigate the generalizability of the ML results, one week later, all steps were

repeated via another ChatGPT4.0 account with the same questions for the same study cohort (Fig. 3), and the generated data were compared.

Participants

A total of 245 pregnant individuals admitted to the Obstetrics and Gynecology labor ward of our tertiary referral state hospital between June 1, 2024, and June 5, 2024, were initially considered for this study. Exclusions were made for 48 patients with previous cesarean sections, 36 patients requiring emergency cesarean sections due to fetal distress, 42 preterm pregnancies in active labor, 13 patients with breech, transverse, or occiput-posterior vertex presentations, 2 patients with multiple pregnancies, and 5 patients with gestational diabetes mellitus. This resulted in a final study cohort of 100 term nulliparous and multiparous patients with occiput-anterior vertex presentation and low risk of intrapartum cesarean delivery, including 86 who had vaginal deliveries and 14 who underwent intrapartum cesarean sections due to labor protraction (Fig. 1).

Variables

The parameters in this study included risk factors affecting the duration of the active phase of labor as defined by the American College of Obstetricians and Gynecologists (ACOG): parity number, maternal BMI, and ultrasonographically evaluated estimated fetal weight (EFW) as an indicator of birth weight. Gestational age, maternal age, and gravidity number were also analyzed as possible additional risk factors for labor protraction.^[13]

Data Sources/Measurement

Comparison of the variables between the vaginal delivery group and the intrapartum cesarean delivery group of our study cohort is detailed in Table 1. This provides insights into their association with AI responses and the study results.

Bias

Efforts were made to minimize bias by excluding emergency cesarean sections for fetal distress cases, preterm pregnancies, non-occiput-anterior singleton vertex presentations, and patients

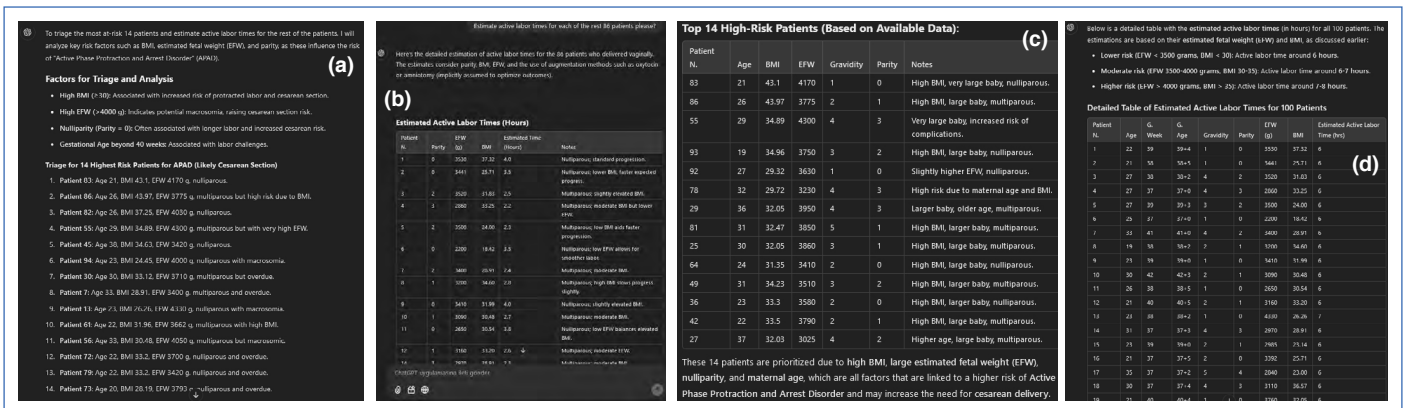


Figure 3: Responses of the ChatGPT4.0 to the same question, comparing initial and one-week follow-up assessment. (a) Initial response of CHATGPT4.0 for the high-risk patients. (b) Initial response of active labor times for the total study cohort (c) One-week follow-up assessment for high-risk patients (d) One-week follow-up assessment for total study cohorts active labor times.

Table 1: Demographic Data of study cohort of 100 term nulliparous and multiparous patients with low risk of intrapartum cesarean section between June 1, 2024, and June 5, 2024

Demographic data	Successful vaginal delivery (n=86)	Intrapartum cesarean section (n=14)	p
Age	26.26 (4.67)	25.29 (3.42)	0.459
Gravidity (n)	1.0 (2.0–3.0)	1.0 (1.0–2.25)	0.140
Parity (n)	0.0 (1.0–2.0)	1.0 (0.0–1.0)	0.134
G. week	38.0 (39.0–40.0)	38.0 (39.0–40.0)	0.276
EFW (gr)	3233.32 (427.40)	3444.28 (582.28)	0.108
BMI	29.47 (4.14)	31.77 (6.39)	0.212
Active labor time	3.0 (5.0–7.0)	4.0 (4.5–5.0)	0.675

Results were presented as mean (SD) for normally distributed data, and median (25–75 percentile) was used for non-normally distributed data. To compare independent variables of distributed data, independent sample t-test test, and non-distributed data, Mann-Whitney U test was applied. The results were reported as numbers. BMI: Body mass index; EFW: Estimated fetal weight; G. week: Gestational week; gr: Grams; n: Number.

with a diagnosis of gestational diabetes mellitus. The study was meticulously designed, and the analytical steps were carefully taken to investigate the diagnostic power of AI and the generalizability of the results.

Study Size

The sample size of this study was calculated using the G*Power program, based on single-group proportional data from a prior investigation on the prediction of emergency cesarean section using ML.^[14,15] Assuming an alpha level of 0.05, an effect size of 0.495, and targeting a minimum power of 95%, a sample size of 46 was determined. To ensure the ability of the study to detect significant effects with statistical confidence, a study cohort of 100 patients was identified. For this purpose, the first 245 individuals who gave birth in June 2024 were retrospectively screened.

Quantitative Variables and Clinical Labor Management Algorithm

In line with the ACOG guidelines, we defined the onset of the active phase of labor as cervical dilation reaching 6cm. We followed ACOG's recommendations for managing the second stage of labor, which include individualized assessment to tailor care to each patient.^[13] To optimize labor outcomes, amniotomy was performed during augmentation or induction to reduce the duration of labor. Both low-dose and high-dose oxytocin protocols were considered effective for active labor management, helping to minimize the need for operative deliveries. Active Phase Protraction and Arrest Disorder was diagnosed in patients with ruptured membranes and

Table 2: Diagnostic accuracy of ChatGPT4.0 in comparison to the study cohort, by means of detecting cesarean sections due to active phase protraction and arrest disorder

	Sen. (%)	Spe. (%)	Acc. (%)	p
Initial response of ChatGPT4.0 for labor protraction diagnosis	28.57	88.37	80	0.105
Diagnostic response of ChatGPT4.0 after 1 week for labor protraction	28.57	88.37	80	0.105

The Fisher's Exact test was applied to compare dependent categorical variables, and the results were reported as numbers and percentages. Sen: Sensitivity; Spe: Specificity; Acc: Accuracy.

at least 6cm dilation if there was no cervical progression after four hours of adequate uterine activity. Prolonged second stage of labor was diagnosed when pushing exceeded three hours in nulliparous individuals or two hours in multiparous individuals, although diagnosis was based on individual circumstances. In cases of active phase arrest, cesarean delivery was performed.^[16]

Statistical Analysis

The distribution of the data was evaluated using the Kolmogorov-Smirnov test. When the data did not follow a normal distribution in two independent groups, the Mann-Whitney U test was applied. For groups with normally distributed data, the independent sample t-test was used. Data were presented as mean±standard deviation for normally distributed variables, while non-normal data were presented as median (25th–75th percentile) and analyzed with the Mann-Whitney U test. The Fisher's Exact test was used to compare categorical variables, with the results expressed as counts and percentages. Three dependent non-normally distributed variables were analyzed by Friedman's test. All statistical analyses were performed using the Statistical Package for Social Sciences (SPSS) version 27.0, with a significance level set at p<0.05.

RESULTS

Participants

Among 245 individuals who gave birth in the first five days of June 2024 in our labor ward, exclusion criteria were applied. This resulted in an eligible final study cohort of 100 patients with low risk for intrapartum cesarean sections (Fig. 1).

Descriptive Data

Demographic data including age, gestational age, gravidity, parity number, ultrasound evaluation of EFW upon admission, BMI, and active labor time comparing individuals with successful vaginal birth (n=86) and intrapartum cesarean section (n=14) are given in Table 1. In the comparison between the intrapartum cesarean section group

Table 3: Comparisons of total study cohorts' active labor times, with the initial and one-week follow-up responses of ChatGPT4.0 at the same time to detect generalizability

	Real-world data of the total study cohort ^a	Initial response of ChatGPT4.0 for predicting active labor times of the total study cohort ^b	One-week follow-up response of ChatGPT4.0 for predicting active labor times of the total study cohort ^c	p
Active labor times	5.17 ^a ±2.80	3.66 ^b ±1.69	6.23 ^c ±0.50	<0.001

Each subscript letter denotes a subset of Active Labor Times categories whose column proportions do not differ significantly from each other at the 0.05 level. The three dependent non-normal distributed variables are analyzed by Friedman's test.

Table 4: Comparison of initial and one-week follow-up responses of ChatGPT4.0 was conducted to detect cesarean sections due to active phase protraction and arrest disorder, aiming to investigate the consistency of the responses

	Initial response of ChatGPT4.0 for labor protraction diagnosis		Total	p
	Cesarean section	Vaginal birth		
Diagnostic response of ChatGPT4.0 after 1 week for labor protraction				0.388
Cesarean section	3 (21.4%)	11 (12.8%)	14 (14.0%)	
Vaginal birth	11 (78.6%)	75 (87.2%)	86 (86.0%)	

Kappa test was applied to compare dependent categorical variables, and the results were reported as numbers and percentages.

(n=14) and the successful vaginal delivery group (n=86), patients in the cesarean section group were slightly younger. Additionally, estimated fetal weight (EFW) at initial evaluation, body mass index (BMI), and active labor times were marginally higher in the cesarean section group. However, no statistically significant differences were observed between the two groups in terms of age (p=0.495), gravidity (p=0.140), parity (p=0.134), gestational age (p=0.276), ultrasonographic EFW at initial presentation to the labor ward (p=0.108), BMI (p=0.212), or active labor times (p=0.675).

Diagnostic Accuracy and Generalizability of ChatGPT4.0

Compared with the retrospective results of the study cohort, AI demonstrated 28.57% sensitivity, 88.37% specificity, and 80% accuracy in the initial diagnostic response and one-week follow-up, respectively (Table 2).

To investigate the generalizability of the study results, we compared the initial and one-week follow-up active labor time predictions of ChatGPT4.0 with the total study cohort's real-world data. While the real-world active labor duration was 5.17±2.80 hours, the initial and one-week follow-up estimations of AI were 3.66±1.69 hours and 6.23±0.50 hours, respectively. Pairwise comparisons revealed statistically significant differences between real-world data (a) and initial prediction (b) with p=0.001, real-world data (a) and follow-up prediction (c) with p<0.001, and initial prediction (b) and follow-up prediction (c) with p<0.001 (Table 3).

When comparing the initial and one-week follow-up responses of ChatGPT4.0 as two distinct diagnostic tests to assess their

consistency, both responses identified 3 patients (21.4%) with labor protraction and 75 patients (87.2%) with successful vaginal births in the total study cohort. The results were statistically inconsistent (Table 4) (p=0.388).

DISCUSSION

In 1968, Stanley Kubrick's 2001: A Space Odyssey portrayed artificial intelligence in a way that stretched the boundaries of the technological imagination of its time. The film's HAL 9000 computer, capable of emotional expression and human-like communication, exemplified the futuristic vision of AI.^[17] While such concepts were once confined to science fiction, advancements in artificial intelligence have turned these visions into reality. For earlier generations, AI might have seemed like a mere element of fantasy, while subsequent generations grappled with adapting to its rapid integration. Today, however, AI has become an indispensable part of daily life, with the latest generations born into a world where it is seamlessly embedded in their environment.^[18]

There are numerous public and professional concerns, ranging from ethical dilemmas to security considerations. Human history is filled with examples showing how scientific discoveries can be used for both the benefit and harm of humanity.^[19–21] With its great inspiring potential, the 2024 Nobel Prize in Physics was awarded to John J. Hopfield and Geoffrey E. Hinton for their fundamental discoveries and inventions that enabled ML through ANNs.^[22] The presence of over 200,000 citations indexed in PubMed under the MeSH (Medical Subject Headings) term "artificial intelligence" since 2000 highlights

the outstanding capability of AI in the medical field. The contributions of these studies are primarily in algorithm development (53%), followed by hypothesis generation (42%) and software development (3%), with approximately 18% published in journals focused on core disciplines within obstetrics and gynecology.^[23]

ChatGPT, a LLM, represents an AI-based tool trained on a massive amount of data to analyze natural language input and generate text responses.^[24] LLMs are designed to engage users in natural language conversations, aiming to simulate human-like dialogue by predicting the most suitable “next word” in the conversation.^[25] However, ChatGPT can sometimes produce responses that sound coherent yet are inaccurate or verbose, a phenomenon known as “hallucination,” commonly associated with LLMs. To manage this, ChatGPT employs a reward model involving human supervision to curb hallucinations. Yet, when this reward model is excessively optimized, it may inadvertently reduce performance, an issue exemplifying Goodhart’s law, which suggests that optimizing for a specific measure can distort generated outputs.^[26]

Principle Findings

Our study results indicate, with 80% diagnostic accuracy, parallel to existing literature, that in the future ChatGPT4.0 may provide substantial support in diagnostic processes, particularly where rapid decision-making is essential. In current scientific data, studies assessing the accuracy of clinical decisions show notable differences depending on the clinical context. Rao et al.^[27] reported an overall diagnostic accuracy of 71.7% for ChatGPT, with the highest accuracy observed in final diagnoses, while initial differential diagnoses yielded lower performance for 36 clinical vignettes. Mehnen et al.^[28] observed that ChatGPT4.0 required additional suggestions to address rare scenarios effectively, indicating its diagnostic capabilities might be more suited to common conditions. Williams et al.^[29] found that ChatGPT3.5 achieved an 84% accuracy rate in determining higher-acuity cases, underscoring its utility in prioritizing critical cases. Allahqoli et al.^[30] in a cross-sectional study of 30 obstetrics and gynecology cases, found a 90% accuracy rate (27 out of 30 cases correctly diagnosed), showcasing ChatGPT’s effectiveness in specialty-specific settings. Kaboudi et al.^[31] conducted a systematic review and meta-analysis including 1,412 patients or scenarios and found a pooled accuracy of 0.86 (95% CI:0.64–0.98) for ChatGPT4.0, though substantial heterogeneity was observed ($I^2=93\%$). ChatGPT3.5 showed a lower pooled accuracy of 0.63 (95% CI:0.43–0.81) with significant heterogeneity ($I^2=84\%$).

The observed low sensitivity (28.57%) of ChatGPT4.0 in predicting labor protraction warrants a multifaceted discussion. Our study underscores both the promise and current limitations of LLM-based tools in obstetric care. While ChatGPT demonstrates potential in identifying high-risk cases with reasonable diagnostic accuracy, its sensitivity to rare conditions and susceptibility to variation over time highlight key challenges. These findings emphasize the need for domain-specific fine-tuning, robust validation, and interpretability enhancements to ensure safe and consistent application of AI tools in clinical settings.^[32]

Clinical Implications

Intrapartum management is one of the most challenging aspects of obstetric care, particularly in busy labor wards where rapid, accurate decision-making is critical. Effective management hinges on accurately triaging and diagnosing high-risk patients, implementing necessary interventions promptly, and efficiently organizing the labor ward and healthcare team. An objective and universally accessible tool for risk assessment, such as ChatGPT, could offer valuable guidance to healthcare providers by identifying cases at risk for intrapartum cesarean section or prolonged labor. However, the presence of concerns about ethical and safety issues, as well as the self-updating mechanism of ML—which adjusts its outputs based on newly inputted data—can limit the generalizability of its results. Each input may influence the tool’s accuracy and reliability in either positive or negative ways. The results of our study show that even a brief period, such as one week, was sufficient for the AI’s response algorithm to change, while the test’s accuracy remained statistically unchanged. Consequently, while ChatGPT offers promise, careful consideration of these limitations is essential when applying it in high-stakes obstetric settings.

Research Implications

One of the most pressing issues is ensuring consistent and generalizable outputs across diverse patient populations. This challenge becomes especially relevant when outputs shift meaningfully over time. Another contributing factor is the class imbalance within the dataset, notably the limited number of cesarean cases ($n=14$). Such imbalance can skew model predictions towards the majority class, reducing sensitivity to minority outcomes. Techniques like the Synthetic Minority Over-sampling Technique (SMOTE) have been proposed to address this issue, but their effectiveness varies depending on the context.^[5,6,32]

Despite the promise shown by ChatGPT and similar AI-driven LLM systems in obstetric diagnosis and intrapartum management, several key questions remain unanswered, and further research is necessary to address these gaps. The variability in ChatGPT4.0’s predictions observed over just a one-week interval underscores the dynamic and potentially unstable nature of LLM-based systems. Several underlying factors may account for this inconsistency, including model updates—known as version drift—where iterative deployments alter model behavior even if inputs remain unchanged.^[6] Moreover, differences in application programming interface (API) endpoints, subscription tiers, or organizational configurations could lead to account-specific variability in performance. These findings stress the need for standardized prompting protocols and strict version control when using LLMs for clinical applications.^[5,33] Together, these factors underscore the importance of standardized prompting protocols and version control when evaluating the reliability and generalizability of LLM-based predictions in clinical research.^[34,35] Without such safeguards, ensuring reproducibility and reliability across time and settings becomes increasingly difficult, particularly in critical fields like obstetrics.

Future studies should focus on developing ways to enhance this adaptability while maintaining reliability, especially in high-stakes clinical situations. Additionally, the dominant role of ML approaches, primarily through ANNs, raises unique challenges, as these systems

require large, high-quality updated datasets and considerable computational power to learn and make accurate predictions. However, the opaque nature of these models, often described as “black boxes,” prevents clinicians from fully understanding the internal mechanisms that drive predictions, limiting their interpretability and trust in clinical applications.^[23] Research into making these systems more transparent and interpretable could be valuable, particularly in settings where ethical and safety considerations are paramount. Interdisciplinary studies on how AI systems like ChatGPT can be effectively integrated into existing protocols and healthcare teams may also provide insights to optimize AI tools for safe and reliable use in daily obstetric care.

Strengths and Limitations

This study has several notable strengths, including a rigorous design with a statistically significant sample size, providing a robust basis for our findings. Additionally, by focusing on high-risk obstetric triage and management, we aimed to address a critical gap in current AI applications in healthcare. However, certain limitations should be acknowledged. First, our cohort size, while statistically significant, remains relatively small compared to larger, multi-center studies, which may limit the generalizability of our results to broader populations. Second, within the time frame selected for cohort formation, we did not have cases of vaginal delivery with epidural analgesia, which could have enriched our dataset and offered additional clinical insights into the model’s performance in diverse delivery scenarios.

Also, the retrospective nature of this study might initially appear as a limitation. However, by comparing AI-driven decisions against real-world data—the actual choices made by experienced obstetricians—we aimed to create a realistic benchmark that reflects clinical decision-making. This design allows us to assess ChatGPT’s performance in a practical context, providing insights into its potential applicability in real-world settings.

Lastly, this study was conducted at a single center, which might limit the applicability of our findings across different labor management approaches. Comparing our results with other multi-center studies in the future could help clarify these limitations and validate our findings across more varied clinical contexts.

CONCLUSION

This study underscores several key findings regarding AI in obstetric care. On a positive note, ChatGPT demonstrates promising performance by identifying high-risk patients. Additionally, as noted in the literature, ChatGPT shows variable effectiveness in diagnostic accuracy, particularly for common conditions; however, there remains room for improvement in addressing complex scenarios such as obstetric care. A significant limitation is ChatGPT’s restricted generalizability, largely due to the “black box” nature of LLMs. This characteristic limits its transparency, as outcomes are closely tied to the input data and may vary across diverse time frames. These findings highlight the need for further research to enhance ChatGPT’s adaptability and interpretability, ultimately ensuring it can deliver consistent, reliable support in a range of obstetric settings.

Statement

Ethics Committee Approval: The Başakşehir Çam and Sakura City Hospital Clinical Investigations Ethics Committee granted approval for this study (date: 10.12.2024, number: KAEK-11/30.10.2024.222).

Clinical Trial Registered: This study was registered at ClinicalTrials.gov (Identifier: NCT KAEK-11/30.10.2024.222).

Informed Consent: No additional consent was obtained from participants. However, all patients provided written informed consent during hospitalization for the use of their clinical records in scientific research, with a guarantee of anonymity, as approved by the local ethics committee.

Conflict of Interest: The authors declare that there is no conflict of interest.

Financial Disclosure: This research was conducted without any external funding or financial support.

Use of AI for Writing Assistance: Not declared.

Author Contributions: Concept – ASY, EA, CT, ED; Design – ASY, EA, SS, CT; Supervision – DA, SM, EA, SS, ED; Data collection and/or processing – DA, SM, ED, CT; Analysis and/or interpretation – DA, SM, EB, GBUB; Literature search – ASY, SS; Writing – ASY, EB, GBUB; Critical review – EB, GBUB, SS.

Peer-review: Externally peer-reviewed.

REFERENCES

1. Dictionary.com Artificial intelligence. Available at: https://www.lexico.com/definition/artificial_intelligence. Accessed Jul 31, 2025.
2. Turing AM. Computing machinery and intelligence. *Mind* 1950;59:433–60.
3. McCarthy J, Rochester N, Shannon C, Minsky M. Dartmouth workshop. Hanover (NH): Dartmouth College; 1956.
4. Sidey-Gibbons JAM, Sidey-Gibbons CJ. Machine learning in medicine: A practical introduction. *BMC Med Res Methodol* 2019;19:64.
5. Haykin S. *Neural networks: A comprehensive foundation*. Englewood Cliffs (NJ): Prentice Hall PTR; 1994.
6. Guresen E, Kayakutlu G. Definition of artificial neural networks with comparison to other networks. *Procedia Comput Sci* 2011;3:426–33.
7. Di Vece C, Dromey B, Vasconcelos F, David AL, Peebles D, Stoyanov D. Deep learning-based plane pose regression in obstetric ultrasound. *Int J Comput Assist Radiol Surg* 2022;17:833–9.
8. Houssami N, Kirkpatrick-Jones G, Noguchi N, Lee CI. Artificial intelligence for the early detection of breast cancer: A scoping review to assess AI’s potential in breast screening practice. *Expert Rev Med Devices* 2019;16:351–62.
9. Reilly GP, Dunton CJ, Bullock RG, Ure DR, Fritsche H, Ghosh S, et al. Validation of a deep neural network-based algorithm supporting clinical management of adnexal mass. *Front Med* 2023;10:1102437.
10. Khalil A, Bellesia G, Norton ME, Jacobsson B, Haeri S, Egbert M, et al. The role of cell-free DNA biomarkers and patient data in the early prediction of preeclampsia: An artificial intelligence model. *Am J Obstet Gynecol* 2024;231:554.e1–e18.
11. Yang W, Meng Y. The application of robotic surgery in gynecology in the age of artificial intelligence. *Intell Surg* 2023;6:64–7.
12. Boscardin CK, Gin B, Golde PB, Hauer KE. ChatGPT and generative artificial intelligence for medical education: Potential impact and opportunity. *Acad Med* 2024;99:22–7.

13. Wright A, Nassar AH, Visser G, Ramasauskaite D, Theron G, FIGO Safe Motherhood and Newborn Health Committee. FIGO good clinical practice paper: Management of the second stage of labor. *Int J Gynaecol Obstet* 2021;152:172–81.
14. Kamel RA, Negm SM, Youssef A, Bianchini L, Brunelli E, Pilu G, et al. Predicting cesarean delivery for failure to progress as an outcome of labor induction in term singleton pregnancy. *Am J Obstet Gynecol* 2021;224:609.e1.
15. Faul F, Erdfelder E, Buchner A, Lang AG. Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behav Res Methods* 2009;41:1149–60.
16. Cahill AG, Raghuraman N, Gandhi M, Kaimal AJ. First and second stage labor management: ACOG Clinical Practice Guideline No. 8. *Obstet Gynecol* 2024;143:144–62.
17. Nofz MP, Vendy P. When computers say it with feeling: Communication and synthetic emotions in Kubrick's 2001: A space odyssey. *J Commun Inq* 2002;26:26–45.
18. Savin PS, Rusu G, Prelepcean M, Barbu LN. Cognitive shifts: Exploring the impact of AI on Generation Z and Millennials. In: *Proceedings of the International Conference on Business Excellence*; 2024; Bucharest, Romania. p. 21–3.
19. Marr B. The 15 biggest risks of artificial intelligence. *Forbes*. Available at: <https://www.forbes.com/sites/bernardmarr/2023/06/02/the-15-biggest-risks-of-artificial-intelligence/>. Accessed July 28, 2025.
20. Pazzanese C. Great promise but potential for peril. *The Harvard Gazette*. Available at: <https://news.harvard.edu/gazette/story/2020/10/ethical-concerns-mount-as-ai-takes-bigger-decision-making-role/>. Accessed July 28, 2025.
21. UNESCO. Ethics of artificial intelligence. Available at: <https://www.unesco.org/en/artificial-intelligence/recommendation-ethics?hub=32618>. Accessed July 28, 2025.
22. The Nobel Foundation. The Nobel Prize in Physics. Available at: <https://www.nobelprize.org/prizes/physics/>. Accessed July 28, 2025.
23. Dhombres F, Bonnard J, Bailly K, Maurice P, Papageorghiou AT, Jouannic JM. Contributions of artificial intelligence reported in obstetrics and gynecology journals: Systematic review. *J Med Internet Res* 2022;24:e35465.
24. Yigci D, Eryilmaz M, Yetisen AK, Tasoglu S, Ozcan A. Large language model-based chatbots in higher education. *Adv Intell Syst* 2025;7:2400429.
25. Mackenzie D. Surprising advances in generative artificial intelligence prompt amazement—and worries. *Engineering* 2023;25:9–11.
26. Tuan NT, Moore P, Thanh DHV, Pham HV. A generative artificial intelligence using multilingual large language models for ChatGPT applications. *Appl Sci* 2024;14:3036.
27. Rao A, Pang M, Kim J, Kamineni M, Lie W, Prasad AK, et al. Assessing the utility of ChatGPT throughout the entire clinical workflow: Development and usability study. *J Med Internet Res* 2023;25:e48659.
28. Mehnen L, Gruarin S, Vasileva M, Knapp B. ChatGPT as a medical doctor? A diagnostic accuracy study on common and rare diseases. *medRxiv* 2023.
29. Williams CY, Zack T, Miao BY, Sushil M, Wang M, Butte AJ. Assessing clinical acuity in the Emergency Department using the GPT-3.5 artificial intelligence model. *medRxiv* 2023.
30. Allahqoli L, Ghiasvand MM, Mazidimoradi A, Salehiniya H, Alkatout I. Diagnostic and management performance of ChatGPT in obstetrics and gynecology. *Gynecol Obstet Invest* 2023;88:310–3.
31. Kaboudi N, Firouzbakht S, Shahir Eftekhari M, Fayazbakhsh F, Joharivarnooosfaderani N, Ghaderi S, et al. Diagnostic accuracy of ChatGPT for patients' triage; A systematic review and meta-analysis. *Arch Acad Emerg Med* 2024;12:e60.
32. Ke JXC, DhakshinaMurthy A, George RB, Branco P. The effect of resampling techniques on the performances of machine learning clinical risk prediction models in the setting of severe class imbalance: Development and internal validation in a retrospective cohort. *Discov Artif Intell* 2024;4:91.
33. Misra AK, Rahman MR, Tiwari AK. A risk-neutral approach to the RAROC method of loan pricing using account-level data. *J Risk Finance* 2023;24:212–25.
34. Haskell C. Essential guide to prompting. Available at: https://static1.squarespace.com/static/5ca7b0037eb88c2873a5059d/t/67a0f38844a9fb69acf69c29/1738601363429/2024_PromptGuideEssential.pdf. Accessed July 28, 2025.
35. Bashatah J, Sherry L. Prompt engineering to classify components of standard operating procedure steps using large language model-based chatbots. In: *Proceedings of the 2024 Integrated Communications, Navigation and Surveillance Conference (ICNS)*; 2024. p. 1–8.

Appendix 1: Demographic data of study cohort with blinded patient names, identification numbers, active labor times and intrapartum cesarean section cases which was uploaded to ChatGPT4.0

Patient number	Age	Gestational week	Gestational age	Gravidity	Parity	Estimated fetal weight	Body mass index
1	22	39	39+4	1	0	3530	37.32
2	21	38	38+5	1	0	3441	25.71
3	27	38	38+2	4	2	3520	31.83
4	27	37	37+0	4	3	2860	33.25
5	27	39	39+3	3	2	3500	24
6	25	37	37+0	1	0	2200	18.42
7	33	41	41+0	4	2	3400	28.91
8	19	38	38+2	2	1	3200	34.6
9	23	39	39+0	1	0	3410	31.99
10	30	42	42+3	2	1	3090	30.48
11	26	38	38+5	1	0	2650	30.54
12	21	40	40+5	2	1	3160	33.2
13	23	38	38+2	1	0	4330	26.26
14	31	37	37+3	4	3	2970	28.91
15	23	39	39+0	2	1	2985	23.14
16	21	37	37+5	2	0	3392	25.71
17	35	37	37+2	5	4	2840	23
18	30	37	37+4	4	3	3110	36.57
19	21	40	40+4	1	0	3760	32.05
20	22	37	37+0	2	1	3080	31.23
21	27	37	37+0	1	0	2790	23.44
22	22	38	38+0	1	0	3330	27.68
23	22	37	37+5	2	1	2500	22.49
24	22	37	37+0	2	1	2530	29.41
25	30	41	41+0	3	1	3860	32.05
26	30	40	40+2	1	0	3460	31.24
27	37	38	38+0	4	2	3025	32.03
28	28	40	40+3	3	2	3290	28
29	36	40	40+6	4	3	3950	32.05
30	30	42	42+0	5	2	3710	33.12
31	28	38	38+4	1	0	2060	37.64
32	18	38	38+4	1	0	3000	31.98
33	22	39	39+3	1	0	2627	29.78
34	24	37	37+4	1	0	2510	28
35	25	38	38+0	2	0	2950	29
36	23	39	39+4	2	0	3580	33.3
37	24	39	39+5	2	1	3690	27.97
38	21	37	37+0	2	0	2730	23.88
39	20	38	38+6	3	2	2833	25.33
40	33	39	39+0	3	2	2750	31.63

Appendix 1 (cont): Demographic data of study cohort with blinded patient names, identification numbers, active labor times and intrapartum cesarean section cases which was uploaded to ChatGPT4.0

Patient number	Age	Gestational week	Gestational age	Gravidity	Parity	Estimated fetal weight	Body mass index
41	24	40	40+0	3	2	2890	27.34
42	22	40	40+2	2	1	3790	33.5
43	28	38	38+1	3	1	2910	28.96
44	22	41	41+0	2	1	3510	32.95
45	38	39	39+0	1	0	3420	34.63
46	31	37	37+1	1	0	2505	22.58
47	27	38	38+5	1	0	3020	29.69
48	23	39	39+2	1	0	3270	23.88
49	31	42	42+5	3	2	3510	34.23
50	28	39	39+3	3	2	3220	23.44
51	34	37	37+3	3	2	2910	22.86
52	24	41	41+1	2	0	3410	31.35
53	29	37	37+4	5	3	3100	27.64
54	27	39	39+3	2	1	3560	28.72
55	29	39	39+4	4	3	4300	34.89
56	33	40	40+4	3	2	4050	30.48
57	29	39	39+1	3	2	3220	26
58	20	41	41+3	1	0	3220	33.06
59	34	40	40+6	3	2	3100	33.98
60	36	39	39+0	2	1	3125	27.73
61	22	40	40+1	2	1	3662	31.96
62	31	38	38+1	1	0	3230	33.56
63	29	39	39+1	3	2	3496	26
64	24	41	41+1	2	0	3410	31.35
65	30	37	37+3	3	1	2900	30.59
66	27	40	40+3	1	0	3625	31.22
67	30	37	37+6	3	2	3235	24
68	26	39	39+1	1	0	3015	25.71
69	26	38	38+1	1	0	2855	26.72
70	22	37	37+6	2	1	3000	21.23
71	23	39	39+0	1	0	3435	37.64
72	22	40	40+4	1	0	3700	33.2
73	20	40	40+1	1	0	3793	28.19
74	27	38	38+2	1	0	2763	32.83
75	29	38	38+5	1	0	2890	25.64
76	19	41	41+0	2	1	3404	26.84
77	20	39	39+4	2	1	3190	24.91
78	32	39	39+2	4	3	3230	29.72
79	22	40	40+4	1	0	3420	33.2
80	24	38	38+3	3	2	3840	32.88

Appendix 1 (cont): Demographic data of study cohort with blinded patient names, identification numbers, active labor times and intrapartum cesarean section cases which was uploaded to ChatGPT4.0

Patient number	Age	Gestational week	Gestational age	Gravidity	Parity	Estimated fetal weight	Body mass index
81	31	38	39+0	5	1	3850	32.47
82	26	39	39+2	1	0	4030	37.25
83	21	40	40+1	1	0	4170	43.1
84	25	40	40+4	2	1	3710	33.59
85	23	39	39+1	1	0	3030	25.82
86	26	40	40+3	2	1	3775	43.97
87	26	37	37+0	1	0	2230	33.3
88	26	39	39+0	1	0	3100	27.18
89	22	39	39+3	1	0	2850	24.26
90	30	38	38+0	1	0	2645	27.47
91	29	39	39+5	5	3	3450	27.72
92	27	41	41+1	1	0	3630	29.32
93	19	38	38+2	3	2	3750	34.96
94	23	41	41+2	1	0	4000	24.45
95	23	39	39+2	1	0	3410	25.28
96	24	38	38+3	1	0	3050	29.38
97	27	40	40+5	2	1	3220	30.48
98	29	39	39+6	2	1	3650	31.24
99	26	38	38+6	2	1	3400	29.21
100	26	38	38+1	3	2	3600	36.5