

The Impact of ChatGPT on Preoperative Anxiety and Postoperative Depression in Cesarean Section: A Prospective Randomized Trial

Background

Cesarean section is frequently associated with significant preoperative anxiety and an increased risk of postpartum depressive symptoms. Digital tools such as ChatGPT may provide accessible, standardized, and personalized information, potentially improving psychological outcomes. However, evidence regarding its effectiveness in this context remains limited. The primary objective of this study is to evaluate the effect of ChatGPT-based information support on preoperative anxiety and postoperative depression in women undergoing cesarean section.

Materials and Methods

Study Design and Setting

This prospective, survey-based study was conducted in two tertiary university hospitals between September 2023 and November 2024. The aim was to evaluate the impact of ChatGPT, used alongside physician consultations, on preoperative anxiety and postoperative depression in pregnant women undergoing elective cesarean section (C/S).

Study Population

Pregnant women scheduled for C/S due to medical indications such as history of uterine surgery, presentation anomaly, multiple pregnancies or placenta previa were included in the study. Patients presenting with active labor, medical complications (e.g., preeclampsia or severe hypertension), or fetal anomalies were excluded. Patients whose newborns required admission to the neonatal intensive care unit were excluded from the study. Participants were randomized in a 1:1 ratio to either a ChatGPT arm or a standard arm using a computer-generated allocation system. Investigators responsible for outcome assessment were blinded to group assignments. Patients scheduled for cesarean delivery due to a history of sphincter injury, active condyloma, vulvar varices, lumbar hernia or other systemic pathologies were classified as having a maternal indication. Placental abnormalities, fetal malposition, intrauterine growth restriction, multiple pregnancies and macrosomia were classified as fetal indication. “Patients with a history of postpartum hemorrhage, intensive care unit admission, neonatal intensive care unit admission, postpartum blood transfusion, retained placenta, postpartum infection, or relaparotomy in a previous pregnancy were evaluated and documented accordingly.

Interventions

Participants in the intervention group were instructed to direct their cesarean section-related questions both to ChatGPT and to their physicians. Patients were allowed to ask clarifying questions to the physician; however, subsequent physician consultations predominantly addressed the same topics raised during the AI-assisted interaction. Each participant was provided access to ChatGPT through a membership system specifically created for this study. The study was conducted at two tertiary public hospitals, with one lead clinician at each site. Participants in the ChatGPT arm were allotted approximately 15–20 minutes to pose up to 10 questions and review the AI-generated responses. Before beginning, they were informed that ChatGPT’s answers might be incomplete or inaccurate. All sessions were supervised by a clinician, and any information judged to be inaccurate or

incomplete was clarified for the participant at the end of the session. Patients could ask up to 10 questions about their surgery, anesthesia, recovery or other related concerns. To ensure confidentiality, the questions and ChatGPT responses were only accessible to the study's principal investigators. No guidance or corrections were provided to ChatGPT regarding the questions asked. However, if ChatGPT's response was not completely accurate, the patient was informed by the physician.

For the control group, participants asked their questions solely to their physicians as part of routine care. The patients in the control group were not subjected to a predefined numerical restriction on the number of questions. Instead, they participated in a routine preoperative consultation with their physician lasting approximately 20 minutes, during which they were free to ask any questions they wished. In our study, observationally, patients in the control group asked a number of questions comparable to the ChatGPT group, with an average of approximately 8–10 questions per consultation.

Both groups asked their questions while hospitalized on the day of their planned cesarean section, prior to surgery. All patients subsequently underwent their scheduled cesarean section according to standard clinical protocols.

Study Procedures

At baseline, all participants completed the State-Trait Anxiety Inventory (STAI) I and II to evaluate anxiety levels and the Edinburgh Postnatal Depression Scale (EPDS) to assess depressive symptoms. The STAI I measures state anxiety (anxiety experienced at the moment), and the STAI II assesses trait anxiety (general tendency to experience anxiety). A score of 40 or above on either scale indicates clinically significant anxiety. For depression, the EPDS was used, where a score of ≥ 10 suggests a high likelihood of depressive symptoms.

After their baseline assessments, participants in the intervention group interacted with ChatGPT following their initial consultations with physicians. Patients were encouraged to ask questions related to surgery, anesthesia, recovery, and newborn care. No restrictions were placed on the complexity or depth of the questions, allowing participants to interact with ChatGPT in a natural manner. All participants were reassessed using the STAI I immediately after the intervention or consultation to evaluate changes in state anxiety. A clinician was present throughout each ChatGPT interaction. Immediately after each patient completed the STAI-I assessment, the clinician reviewed the ChatGPT responses for potential inaccuracies or incomplete information and provided clarifications as needed. In addition to this real-time review, a formal accuracy evaluation of all ChatGPT responses was conducted after the study was completed. Postoperative depressive symptoms were reassessed one week after C/S using the EPDS.

Visual pain scores (VPS) were also recorded postoperatively to evaluate pain levels. Pain intensity was assessed using a 0–10 numeric rating scale, where 0 indicated no pain and 10 indicated the worst imaginable pain. VPS measurements were collected during routine follow-up in the first week postpartum. The Turkish validated versions of the Edinburgh Postnatal Depression Scale (EPDS) and the State-Trait Anxiety Inventory (STAI) were used in this study

ChatGPT Response Validation

After the study was completed and the target sample size was reached, the responses provided by ChatGPT were reviewed for accuracy and reliability. Responses were categorized into three groups: "appropriate," "inappropriate," and "unreliable." "Appropriate" responses fully adhered to existing clinical guidelines or evidence-based information, providing accurate and relevant content. "Inappropriate" responses included those with incomplete information, medically incorrect suggestions, or content disconnected from the clinical context. "Unreliable" responses were characterized by a lack of verifiable medical information or ambiguous content that could potentially mislead users. The evaluation was conducted independently by two obstetricians (B.M.S. and K.H.) in a double-blind manner. Discrepancies in evaluation were resolved through mutual discussion to ensure consistency. This analysis aimed to evaluate the quality of information provided by ChatGPT and its alignment with clinical standards, using authoritative guidelines such as those from the World Health Organization (WHO) and the National Institutes of Health (NIH) as references.

All participants in the ChatGPT group were subsequently counseled by the attending physician, during which any incomplete, inaccurate, or potentially misleading information generated by ChatGPT was reviewed and corrected to ensure patient safety and adherence to clinical standards.

Statistical Analysis

A priori power analysis was conducted to determine the required sample size using G*Power software version 3.1. Based on prior literature and clinical relevance, a minimum clinically important difference of 4 points in the post-intervention STAI-I scores between groups was assumed, with a standard deviation of 10. An effect size (Cohen's *d*) of 0.4 was calculated accordingly. To achieve a power of 90% at a two-sided alpha level of 0.05, the required sample size was estimated to be 133 participants per group. Considering a potential attrition rate of 10–15%, the target enrollment was increased to 150 participants per group, resulting in a total sample size of 300 participants.

Demographic information, anxiety and depression scores, vital signs, and visual pain scores were collected for all participants. ChatGPT interactions, including the questions asked and responses received, were securely stored to maintain data privacy. Statistical analysis was conducted using IBM SPSS Statistics version 23. Descriptive statistics summarized the data, while inferential statistics, including the Mann-Whitney U test for continuous variables and Pearson's chi-square test for categorical variables, were employed to compare the groups. Normality was assessed using the Shapiro-Wilk and Kolmogorov-Smirnov tests, and statistical significance was defined as $p < 0.05$.

Ethical Considerations

Ethical approval for the study was obtained from the Lokman Hekim University Ethics Committee (2023/9 – No: 2023158). All participants provided written informed consent prior to enrollment. Data privacy and confidentiality were strictly maintained, with only the principal investigators having access to participants' data, including ChatGPT interactions.