OPEN ACCESS

**Saudi Journal of Medical and Pharmaceutical Sciences**
Abbreviated Key Title: Saudi J Med Pharm Sci
ISSN 2413-4929 (Print) | ISSN 2413-4910 (Online)
Scholars Middle East Publishers, Dubai, United Arab Emirates
Journal homepage: https://saudijournals.com

**Original Research Article** **Psychology**

# Evaluating the Effectiveness of a Large Language Model in the Psychological Assessment of Potential Liver Transplant Candidates: A Feasibility Study

Wadha A.E. Alqahtani, MSc[1,2], Dimitri A. Raptis, MD, MSc, PhD[1], Dieter C. Broering, MD, PhD, FEBS, FACS[1,2], Mamdouh Alenazi, BSc, MSc, PhD[1,2*]

[1]Organ Transplant Cetner of Excellence, King Faisal Specialist Hospital and Research Center, Riyadh, Saudi Arabia
[2]Master of Clinical Psychology, College of Medicine, Alfaisal University, Riyadh, Saudi Arabia

**DOI:** https://doi.org/10.36348/sjmps.2024.v10i10.006 | **Received:** 02.09.2024 | **Accepted:** 09.10.2024 | **Published:** 22.10.2024

***Corresponding author:** Dr Mamdouh Alenazi
1Organ Transplant Cetner of Excellence, King Faisal Specialist Hospital and Research Center, Riyadh, Saudi Arabia

## Abstract

***Introduction*:** The use of artificial intelligence (AI) in medical settings has shown promise in various domains including diagnostics, treatment recommendation, and patient management. Recent advances have explored the potential of AI in pre-surgical assessments, but its application in transplant psychology remains unexplored. The objective of this study is to assess the effectiveness of AI in the context of screening potential liver transplant donors and recipients. To assess the feasibility of using ChatGPT-4 to screen potential liver transplant donors and recipients. ***Methods*:** This study utilizes a cross-sectional research design to evaluate the feasibility of using ChatGPT-4 in the preliminary screening of living liver donors and liver transplant candidates. The study aims to determine the accuracy and reliability of ChatGPT-4 in assessing hypothetical scenarios involving potential donors and recipients. ***Results*:** The analysis showed no found no significant differences between ChatGPT-4 and the expert panel in assessing liver transplant candidates, demonstrating an overall accuracy of 83.58%, sensitivity of 56.10%, and specificity of 80.49%. Additionally, the Cohen's Kappa statistic of 0.68 (95% CI: 0.52-0.83) indicated substantial agreement between ChatGPT-4 and the psychologists' evaluations. The absence of false positives (0%) and a low false negative rate (8%) emphasize ChatGPT-4's cautious and accurate decision-making capabilities. ***Conclusion*:** The findings of this study demonstrate that ChatGPT-4 has the potential to serve as an effective screening tool for liver transplant candidates, complementing the work of human experts and enhancing the overall efficiency of the transplant process. While challenges remain, the integration of AI into the liver transplantation workflow could lead to significant improvements in candidate evaluation and patient outcomes, paving the way for the broader application of AI in clinical practice.

**Keywords:** Large language model; potential liver transplant candidates; Psychological assessment; Feasibility study.



## INTRODUCTION

Liver transplantation is a highly effective, life-saving procedure for patients with end-stage liver disease, extending life expectancy by approximately 15 years [1]. It crucially depends on the meticulous selection of candidates by a multidisciplinary team [2]. A key part of this selection is the psychosocial evaluation, which assesses potential psychological issues and psychiatric comorbidities that could affect outcomes [2]. This includes evaluating mental health, adherence to treatment, substance use history, and the availability of caregivers and social support, especially for patients with encephalopathy [1].

Psychosocial assessments also help manage challenges faced by candidates, including adapting to chronic diseases and the stress of the waiting list, which can lead to significant psychosocial complications affecting both patients and their families [2]. The Organ Procurement and Transplantation Network (OPTN) provides comprehensive guidelines for these evaluations, emphasizing the need for thorough psychological, social, and informed consent assessments to ensure donor suitability and safety [3].

However, the screening process faces challenges like time constraints, resource limitations, and potential biases, calling for more resources, ongoing research, and efforts to uphold ethical standards [3]. AI

is increasingly used to enhance disease diagnosis and treatment personalization, demonstrating significant potential in medical diagnostics and patient safety by analyzing complex medical data and providing evidence-based treatment recommendations [4, 5].

Large Language Models such as ChatGPT-4, Gemini, and LLaMA, utilize advanced machine learning for tasks such as text classification and summarization. While they offer promising advancements in healthcare, their application in liver transplant candidate screening remains under-researched [6].

This study aims to assess the feasibility of using ChatGPT-4 for this purpose, comparing its performance with human experts in evaluating candidates based on various scenarios. This may enhance screening efficiency, reduce bias, and inform future AI developments in transplantation and beyond.

## METHODS

### Research Design

This study used a cross-sectional research design to evaluate the feasibility of using ChatGPT-4 (Open AI, San Francisco, USA) in the preliminary screening of living liver donors as well as liver transplant candidates. The study aimed to determine the accuracy and reliability of ChatGPT-4 in assessing hypothetical scenarios involving potential donors and recipients.

### Data Collection

The study evaluated a total of 60 hypothetical profiles divided equally between potential living liver donors (30 profiles) and liver transplant candidates (30 profiles). The hypothetical profiles were generated randomly by specialist transplant psychologists (MA), ensuring a diverse representation of demographics and medical histories. This diversity was crucial to evaluate the robustness of ChatGPT-4 across various scenarios. Since the study involved hypothetical scenarios, there was no recruitment of real subjects. All data used were artificial and generated for the purpose of this research. All possible generated profiles were included in the study to ensure a broad and comprehensive assessment. There were no specific exclusion criteria as the focus was on evaluating the feasibility across a wide spectrum of cases.

### Experimental Procedure

The experiments in this study were structured in 8 weeks, from 1-6-2024 to 25-7-2024.

In Figure 1, there is a detailed flowchart representing the study process, from scenario development to the final conclusion. Starting with the development phase of hypothetical scenarios during the first two weeks. This initial stage involved setting up and calibrating the hypothetical scenarios. The evaluation phase was conducted over the next four weeks, where hypothetical profiles were inputted into ChatGPT-4, and their evaluations were documented. Each profile was evaluated for their suitability as a donor or recipient, determine whether they are psychologically clear for donation/transplantation, requiring further assessment, or not psychologically clear for donation/transplantation. The last two weeks, weeks seven and eight, were spent on data analysis and report preparation. During this time, the evaluations that were documented were analyzed, and the findings were compiled into the final report.

### Data Analysis

The performance of ChatGPT-4 was evaluated against expert assessments using several metrics. Sensitivity measured ChatGPT-4's ability to accurately identify candidates who were either clear or not clear for liver transplantation. Specificity it assessed its capability to correctly exclude those who were not suitable for the procedure. The Area Under the ROC Curve (AUC) provided a comprehensive measure of the model's overall diagnostic performance. Absolute agreement evaluated the degree of concordance between the decisions made by ChatGPT-4 and human experts, measuring the percentage of cases in which both the model and the experts agree on a candidate's suitability for liver transplantation. Additionally, the Kappa agreement utilized the Kappa statistic to measure inter-rater agreement between ChatGPT-4 and human experts, accounting for the chance occurrence of agreement and offering a robust assessment of the model's reliability compared to expert judgment. Statistical analysis will be conducted using statistical analysis was performed using R version 3.3.2 (R Core Team, GNU GPL v2 License), R Studio version 1.0.44 (RStudio, Inc. GNU Affero General Public License v3, Boston, MA, 2016) with the graphical user interface (GUI) rBiostatistics.com alpha version (rBiostatistics.com, Riyadh, KSA, 2023) [7].

## RESULTS

In this study, a total of 60 hypothetical scenarios, evenly split between donors and recipients, were created to provide a wide range of demographic variables such as age, gender, and marital status. These scenarios were designed to include diverse demographic profiles, family and social histories, and psychological and psychiatric symptoms, presenting complex cases with multiple comorbidities and varying levels of social support. The aim was to comprehensively assess ChatGPT-4's evaluation capabilities.

ChatGPT-4 assessed these candidates, categorizing 28 as cleared, 11 for further assessment, and 21 as not cleared (Table 1). Conversely, the expert panel of transplant psychologists evaluated the same profiles, distributing their assessments evenly across the three categories: 20 for further assessment, 20 cleared, and 20 not cleared (Table 1). The concordance rates between ChatGPT-4 and the expert panel were 71.4% for clearance, 81.8% for further assessment, and 85.7% for non-clearance (Figure 2). The Cohen's Kappa statistic

indicated substantial agreement at 0.68, with a p-value of <0.001, suggesting statistically significant alignment.

The accuracy metrics revealed a sensitivity of 86.66%, specificity of 90%, and both positive and negative predictive values at 88.24% and 85.71%, respectively. The confusion matrix analysis (Table 2) detailed the distribution of true positives (32), false negatives (8), and true negatives (20), with no false positives recorded, indicating strong performance and careful decision-making by ChatGPT-4.

The ROC curve, with an Area Under the Curve (AUC) of 0.718, highlighted ChatGPT-4's effective distinction between cleared and non-cleared candidates, suggesting a high true positive rate and a low false positive rate (Figure 1). Notably, the presence of 8 false negatives suggested a conservative approach by ChatGPT-4, especially in complex or borderline cases, while the absence of false positives underscored its accuracy.

Statistical tests found no significant differences between ChatGPT-4 and the expert panel in assessing liver transplant candidates, demonstrating an overall accuracy of 83.58%, sensitivity of 56.10%, and specificity of 80.49%. The results confirmed substantial agreement between ChatGPT-4 and the psychologists, as evidenced by a kappa agreement of 0.68 for overall evaluations and 0.73 for clearance-specific assessments, both with a p-value of <0.001, indicating robust alignment and consistent decision-making between the AI model and human experts.

**Table 1: ChatGPT-4 assessed the 60 candidates and categorized them as follows**

| Assessment Type | Frequency | Percentage |
|---|---|---|
| Assessment | 11 | 18.33% |
| Cleared | 28 | 46.67% |
| Not Cleared | 21 | 35.00% |
| **Total** | **60** | **100%** |

**Table 2: The expert of transplant psychologists evaluation, consisting of two transplant psychologists, categorized the same 60 scenarios as follows**

| Assessment Type | Frequency | Percentage |
|---|---|---|
| Assessment | 20 | 33.33% |
| Cleared | 20 | 33.33% |
| Not Cleared | 20 | 33.33% |
| **Total** | **60** | **100%** |

**Table 3: Confusion Matrix Analysis: true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN)**

| | Predicted Cleared | Predicted Not Cleared |
|---|---|---|
| **Cleared** | TP: 32 | FN: 8 |
| **Not Cleared** | FP: 0 | TN: 20 |

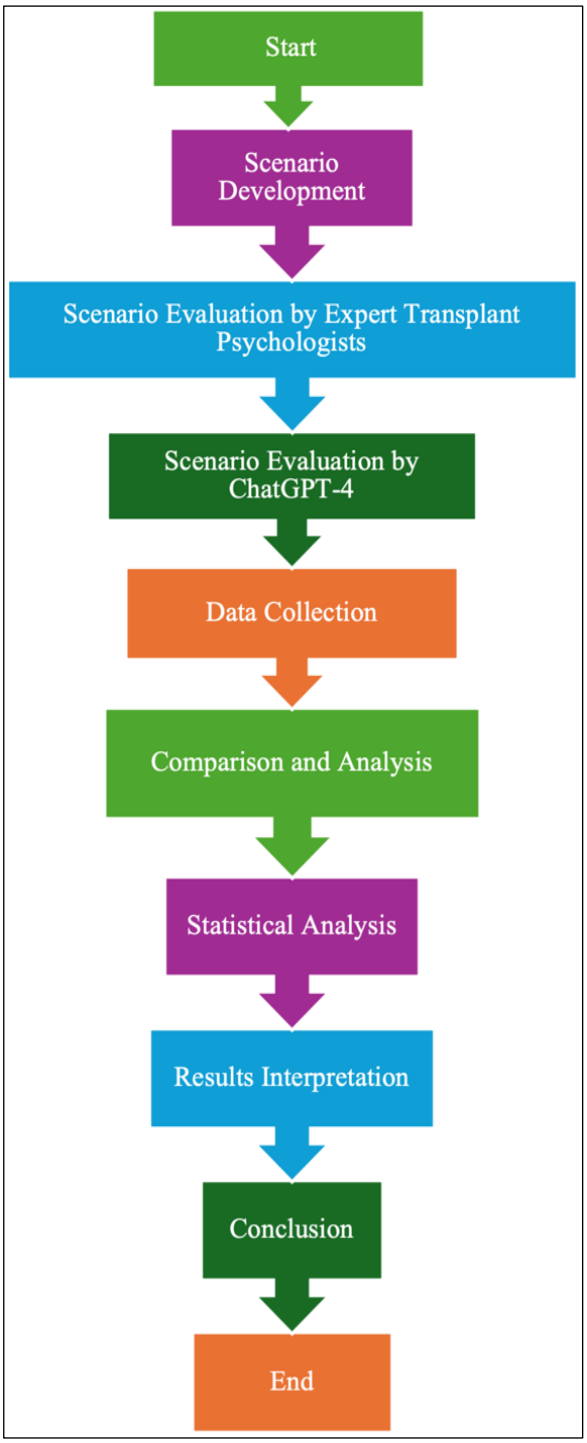


**Figure 1: Comprehensive flowchart demonstrating the study process from scenario development to the conclusion**

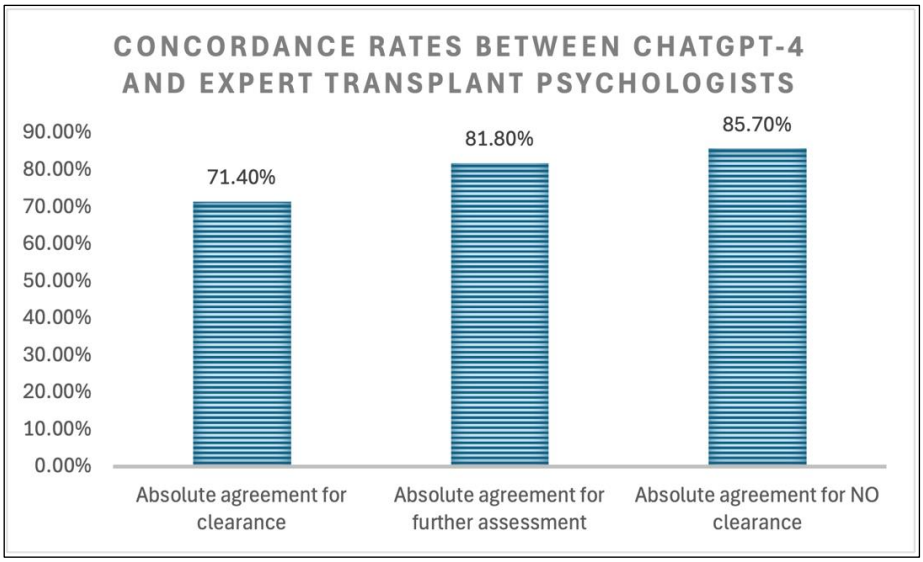


**Figure 2: Concordance rates between ChatGPT-4 and Expert Transplant Psychologists**

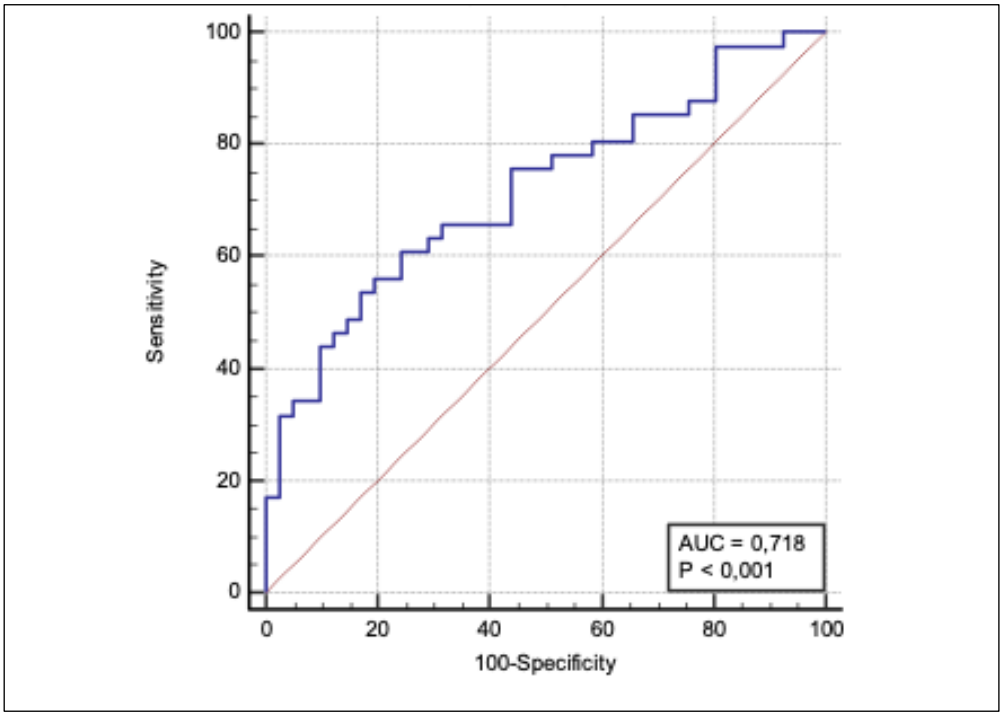


**Figure 3: ROC curve. The Area Under the Curve (AUC) is 0.718, which indicates strong performance of ChatGPT-4 in distinguishing between cleared and non-cleared candidates according to the expert panel's evaluations. This curve demonstrates that the model performs well, with a high true positive rate (sensitivity) and a low false positive rate**

## DISCUSSION

Recent advances in artificial intelligence (AI) technology have led to the development of large language models like ChatGPT, which have shown potential in various fields, including healthcare [8]. Further studies demonstrate the effectiveness of ChatGPT in clinical decision support, patient consultation, and medical documentation [9]. For instance, research has shown that ChatGPT can assist in clinical decision-making by providing up-to-date medical research and guidelines, thereby aiding doctors in making more informed decisions [9]. Additionally, ChatGPT has been found useful in generating medical documents, such as discharge summaries and pathology reports, which can alleviate the documentation burden on clinicians and enhance the quality of patient care [9].

This study aimed to assess the feasibility of utilizing ChatGPT-4 as a screening tool for liver transplant candidates, comparing its performance with that of expert transplant psychologists. The findings indicate substantial agreement between ChatGPT-4 and the expert panel, suggesting that ChatGPT-4's evaluations align closely with those of experienced transplant psychologists. The absence of false positives and a relatively low rate of false negatives underscore

ChatGPT-4's reliability in screening, suggesting its potential utility as an adjunct to human evaluators to reduce clinicians' workload and enhance consistency in assessments.

ChatGPT-4's conservative approach in handling complex or borderline cases—indicated by a higher incidence of false negatives compared to false positives—suggests a prioritization of caution, a critical attribute in medical settings. This approach may help mitigate the risk of adverse post-transplant outcomes by ensuring only suitable candidates advance in the evaluation process. Furthermore, the specificity of 80.49% and sensitivity of 56.10%, along with an AUC of 0.718, confirm ChatGPT-4's effectiveness in identifying suitable candidates and excluding inappropriate ones, crucial for successful liver transplantation.

Integrating ChatGPT-4 into the liver transplantation screening process could profoundly impact clinical practice by significantly reducing the time and resources required for preliminary evaluations. This would allow transplant teams to concentrate more on intricate cases needing thorough analysis. Additionally, by minimizing human biases related to socioeconomic status, race, or gender, ChatGPT-4 could help achieve a more equitable and consistent screening process. The model's capacity to rapidly process and analyze extensive data sets could also improve the efficiency of transplant programs, potentially shortening the waiting time for candidates, thereby enhancing patient outcomes.

Despite promising results, the study acknowledges limitations, such as the potential for bias in ChatGPT-4's training datasets, which could lead to biased or misleading outcomes. Ethical considerations also persist regarding AI's role in sensitive medical decision-making. The study supports using AI as a supplementary tool, not a replacement for human judgment, emphasizing that human experts should make final decisions. Another limitation is that the hypothetical scenarios used in this study were designed to closely match real-life cases encountered in clinical practice. Despite the fact that this approach provides controlled analysis, it may not capture all the psychological and emotional complexities of actual cases. Recommending that future research incorporates real case studies to gain a better understanding of the nuanced human factors involved.

This research prompts further investigation into integrating ChatGPT-4 with other AI tools to develop a more comprehensive screening system. Additional studies are needed to evaluate the model’s performance across diverse populations and its long-term impact on transplant outcomes. As AI technology evolves, future research should aim to refine models like ChatGPT-4 to address identified limitations and develop guidelines for the ethical use of AI in clinical settings, ensuring responsible and transparent application.

## CONCLUSION

ChatGPT-4 demonstrates significant potential as an effective screening tool for liver transplant candidates, enhancing the efficiency of the transplant process and complementing human expertise. While challenges remain, the integration of AI into the transplantation workflow could lead to substantial improvements in candidate evaluation and patient outcomes, setting the stage for broader AI application in clinical practice.

## REFERENCES

1. Yara Dababneh & Omar Y. Mousa. Liver Transplantation. (StatPearls, 2023).
2. García-Alanís, M., Toapanta-Yanchapaxi, L., Vilatobá, M., Cruz-Martínez, R., Contreras, A. G., López-Yáñez, S., ... & García-Juárez, I. (2021). Psychosocial evaluation for liver transplantation: A brief guide for gastroenterologists. Revista de Gastroenterología de México (English Edition), 86(2), 172-187.
3. OPTN. Guidance for the Medical Evaluation of Potential Living Liver Donors. https://optn.transplant.hrsa.gov/professionals/by-topic/guidance/guidance-for-the-medical-evaluation-of-potential-living-liver-donors/ (n.d.).
4. Alowais, S. A., Alghamdi, S. S., Alsuhebany, N., Alqahtani, T., Alshaya, A. I., Almohareb, S. N., ... & Albekairy, A. M. (2023). Revolutionizing healthcare: the role of artificial intelligence in clinical practice. *BMC medical education*, *23*(1), 689. doi: https://doi.org/10.1186/s12909-023-04698-z.
5. Junaid, B., Usman, M., Aditya, N., & Bryan, W. (2021). Artificial intelligence in healthcare: transforming the practice of medicine. *Future Healthc. J.*, *8*, e188–e194.
6. Li, J., Dada, A., Puladi, B., Kleesiek, J., & Egger, J. (2024). ChatGPT in healthcare: A taxonomy and systematic review. *Comput. Methods Programs Biomed.*, *245*, 108013.
7. rBiostatistics.com. rBiostatistics.com: A Cloud-Based Graphical User Interface for R Statistics and an eLearning Platform. (2017).
8. Alanezi, F. (2024). Assessing the Effectiveness of ChatGPT in Delivering Mental Health Support: A Qualitative Study. *J. Multidiscip. Healthc*, 461–471. doi: https://doi.org/10.2147/JMDH.S447368.
9. Mu, Y., & He, D. (2024). The Potential Applications and Challenges of ChatGPT in the Medical Field. *International Journal of General Medicine*, 817-826. doi: https://doi.org/10.2147/IJGM.S456659.