Evaluating ChatGPT's Responses to Commonly Asked Questions in Allogeneic Hematopoietic Stem Cell Transplantation: A Physician-Based Assessment

Safa S Afridi[1], Sumbal Aziz[2], Muhammad Junaid Tariq[3], Showkat Hamid[3], Muhammad Salman Faisal[3], Jonathan Harrison[1],Mulham Ombada[3], Anuja V. Abhyankar[3], Faiz Anwer[4], Raheel Iftikhar[5], Shernan Holtan[3], Mehdi H. Hamadani[6], Zunairah Shah[3]

1. SUNY Upstate Medical University, Syracuse, NY, USA
2. Advent Health Sebring, FL, USA
3. Roswell Park Cancer Center, Buffalo, NY, USA
4. Cleveland Clinic, Taussig Cancer Institute, Cleveland, OH, USA
5. Armed Forces BMT Center, NUMS, Rawalpindi, Pakistan
6. Froedtert & Medical College of Wisconsin, Milwaukee, WI, USA

Introduction

Artificial intelligence (AI) tools such as ChatGPT are increasingly used by patients to understand complex treatments like allogeneic hematopoietic stem cell transplantation (HSCT). However, the clinical accuracy and usefulness of AI-generated responses in this context remain underexplored.

Aim

To evaluate the quality and clinical accuracy of ChatGPT's responses to commonly asked questions about allogeneic hematopoietic stem cell transplantation (HSCT).

Objectives

- Collect frequently asked HSCT questions from CDC resources, patient forums, and clinical practice.
- Assess ChatGPT's responses to these questions using the Global Quality Scale (GQS).
- Identify strengths and limitations in ChatGPT's performance, with focus on patient education relevance.

Methods

•**Design:** Cross-sectional evaluation of ChatGPT (GPT-4).

•**Questions:** Sixteen frequently asked HSCT questions sourced from CDC educational materials, online forums, and clinical practice.

•**Responses:** Generated using ChatGPT and compiled in a survey format.

•**Reviewers:** Thirteen transplant physicians (directors and academic faculty).

•**Assessment:** Each response rated on the 5-point **Global Quality Scale (GQS)**
(1 = very poor, 5 = excellent).

•**Analysis:** Mean scores and standard deviations (SD) calculated for each question.

Discussion

- ChatGPT responses are **readable and empathetic**, but clinical depth is insufficient for high-stakes topics.
- Risk of **oversimplification + missing context;** may mislead patients if not supervised.
- AI has **potential for educational support** but must not replace physician guidance.

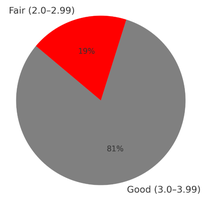
Results

The **overall mean Global Quality Scale (GQS) score** across 16 questions was **3.43 ± 0.93**, reflecting **moderate quality**.

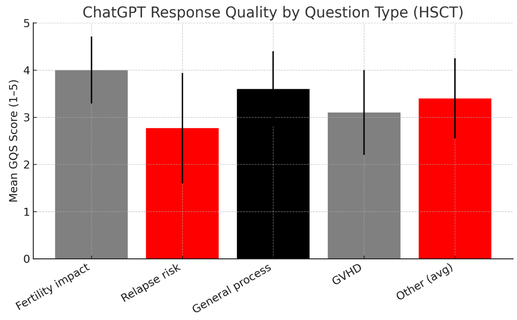
- **Highest-rated response**: fertility-related impact (**4.0 ± 0.71**).
- **Lowest-rated response**: relapse risk (**2.77 ± 1.17**).
- **Rating distribution**:
  81% (13/16) classified as *Good* (GQS 3.0–3.99).
  19% (3/16) classified as *Fair* (GQS 2.0–2.99).

Patterns observed:

- General, process-focused questions were rated higher.
- Responses to complex issues (e.g., GVHD, relapse) lacked nuance and depth.
- Reviewers valued the accessible, empathetic tone but highlighted oversimplification and missing clinical details.



HSCT Question	Mean GQS	SD
Fertility impact	4.0	0.71
Relapse risk	2.77	1.17
General process	3.6	0.8
GVHD	3.1	0.9
Other (avg)	3.4	0.85



Conclusion

- ChatGPT shows **moderate utility** in HSCT education.
- Best use: simple patient FAQs.
- Weakness: nuanced/complex issues (GVHD, relapse).
- Patients must be **educated about ChatGPT's fallacies** (outdated info, lack of personalization).
- Recommendation: oncology-specific training and expert oversight.

Acknowledgements

We thank all transplant experts and faculty who participated.

© 2025 SUNY Upstate Medical University and Collaborating Institutions.
Presented at SOHO 2025 Annual Meeting, Houston, TX.