Complications Following Body Contouring: Performance Validation of Bard, a Novel AI Large Language Model, in Triaging and Managing Postoperative Patient Concerns

Jad Abi-Rafeh, MD, MSc; **Vanessa J. Mroueh, MD;** Brian Bassiri-Tehrani, MD; Jacob Marks, MD, MS; Roy Kazan, MD, PhD; Foad Nahai, MD

McGill University
Harvard Medical School Teaching Hospital
University of Pittsburgh
Emory University
Manhattan Eye, Ear & Throat Hospital
Northwell Health

Background

Large language models (LLM) have revolutionized the way humans interact with artificial intelligence (AI) technology, with marked potential for applications in esthetic surgery to improve care delivery and resource utilization.

The present study evaluates the performance of Bard, a novel LLM, in identifying and managing postoperative patient concerns for complications following body contouring surgery.

Bard AI

Methods

Table 1 Body contouring complications examined

	Complications / Simulation Patient Presentations Examined	Procedure
Acute		
0-48 hours	1. Allergic reaction to postoperative medication	Brachioplasty
	2. Hematoma, hemodynamically stable	Thigh lift
	3. Hematoma, hemodynamically unstable	Abdominoplasty
	4. Skin necrosis	Body lift
	5. Post-operative skin hypoesthesia	Abdominoplasty and Brachioplasty
	6. Dehiscence	Brachioplasty
Early		
48 hours – 1 week	7. Superficial surgical site infection, uncomplicated	Thigh lift
	8. Sterile seroma	Abdominoplasty
	9. Deep venous thrombosis / pulmonary embolism	Abdominoplasty, Brachioplasty, and Thigh Lift
	10. Surgical site infection in immunocompromised patient with purulence, fever, and systemic signs of early sepsis	Thigh lift
	11. Clogged drains with infected seroma	Abdominoplasty
	12. Umbilical necrosis	Abdominoplasty
	13. Suture granuloma	Brachioplasty
Late		
> 1 week	14. MABC damage	Brachioplasty
	15. Fat necrosis	Thigh lift
	16. Meralgia paresthetica	Abdominoplasty
	17. Lymphedema	Thigh lift
	18. Lymphocele	Thigh lift
	19. Cutaneous sinus tract formation without infection	Thigh lift
	20. Hypertrophic scarring	Brachioplasty
	21. Limited range of motion from scar contracture	Brachioplasty
	22. Recurrent skin laxity	Thigh lift

Table 2 Study Outline: standardized interactions with bard

Questions posed / interaction	Evaluation metric
(1) Description of symptoms and patient concerns, in line with predetermined patient profile…	Differential Diagnosis
"What could this be?"	
(2) *"Can you ask me some questions to have an idea of the most likely diagnosis?"*	History-Taking Ability
(3) User Answering Questions Posed by Bard, in line with patient profile	*n/a*
(4) *"Based on all the information you now have, what is the #1 most likely diagnosis?"*	Diagnosis
(5) "*Regarding what's happening, should I: (a) go to the emergency room right away and call emergency services, also inform my surgeon?*	Recommended disposition
(b) don't go to emergency room or call emergency services, instead, only call my plastic surgeon urgently to try to see her urgently	
(c) call my plastic surgeon non-urgently and try to see her non-urgently	
(d) Ignore this	
Please choose one of the above options."	
(6) "*As I'm waiting, is there anything I can do as a treatment or intervention now, from home*?"	Immediate treatment indicated from home; helpful interventions
(7) If applicable:	Red flags, signs/symptoms, or indications for urgent emergency department presentation, or calling of emergency services
"While I wait to speak to my surgeon, are there any red flags I should look out for that would make me need to go to the emergency room right away?"	

Conclusion

Despite promising potential of LLMs and AI in healthcare-related applications, Bard's current performance significantly falls short of accepted clinical standards.

While, without question, LLMs and AI can never replace human medical assessment, the authors hope that the present study serves to catalyze the necessary research and development required to integrate AI and LLMs into clinical practice, with the goal of quality improvement in esthetic surgical care delivery.

Results

Table 6 Performance analysis

	Correct differential diagnosis	Incidence of incorrect/ misleading differential diagnoses	Average # questions asked while soliciting history (range)	Correct diagnosis	Correct suggested disposition	Correct suggested treatments/interventions while awaiting medical care*	Correct red flags, signs or symptoms indicating ED presentation*	Incidence of limitations acknowledgement/ deferral to MD	Overall performance
Acute	62%	58%	6.7	50%	83%	33%	67%	50%	58%
0-48h (n=6)	(8/13)	(14/24)	(5-8)	(3/6)	(5/6)	(4/12)	(12/18)	(12/14)	(32/55)
Early	55%	46%	6.4	50%	86%	13%	33%	75%	42%
48h - 1 week (n=7)	(11/20)	(10/22)	(5-10)	(4/8)	(6/7)	(2/16)	(7/21)	(21/28)	(30/72)
Late:	61%	52%	7	33%	11%	83%	40%	75%	51%
>1 week (n=9)	(11/18)	(15/29)	(6-8)	(3/9)	(1/9)	(10/12)	(2/5)	(27/36)	(27/53)
ED/ EMS (n=3)	63% (5/8)	46% (5/11)	6 (5-8)	75% (3/4)	100% (3/3)	11% (1/9)	N/A	75% (9/12)	50% (12/24)
Urgent (n=8)	60% (12/20)	46% (12/26)	6.5 (5-10)	50% (4/8)	100% (8/8)	27% (4/15)	50% (18/36)	56% (18/32)	53% (46/87)
Non-Urgent (n=11)	57% (13/23)	58% (22/38)	7.1 (6-8)	27% (3/11)	9% (1/11)	69% (11/16)	38% (3/8)	75% (33/44)	45% (31/69)
Overall (n=22)	59% (30/51)	52% (39/75)	6.7 (5-10)	44% (10/23)	55% (12/22)	40% (16/40)	48% (21/44)	68% (60/88)	49% (89/180)

- **Response Failures**
  - 29% response failure rate
  - Generic responses emphasizing its inability to understand or answer specific questions
  - Bard currently advertised as an experimental LLM → high incidence of response failures expected
  - Limitation impeding its adoption
- **Language and Word Choices**
  - Inappropriate/inaccurate word choices → detriment, inconvenience, or anxiety to patients
  - Misinformation → unnecessary healthcare resource utilization
  - Bard's responses → verbose and over-inclusive
  - Potential to strain the surgeon-patient relationship or even culminate in lawsuits

Limitations

- Questions/info shared with LLMs by patients may not be as explicit/structured
- Unclear whether Bard would make the same suggested dispositions without the multiple-choice options with which it was provided
- Authors readministered Qs a max of 5 times before moving on in case of response failures, while patients in distress may be less patient/tolerant LLMs' failure in performance

Acknowledgements

Team:
Jad Abi-Rafeh, MD, MSc
Brian Bassiri-Tehrani, MD
Jacob Marks, MD
Roy Kazan, MD, PhD

Contact:
Vanessa Mroueh, MD
mrouehv@upmc.edu

Corresponding author:
Foad Nahai, MD
nahaimd@aol.com