APCN x TSN 2025
23rd Asian Pacific Congress of Nephrology
Gene, Immunology, Vast, MEtabolism at its Finest!

Evaluating Generative Artificial Intelligence (AI) Models for Patient Education on Immunosuppression Post-Kidney Transplantation:

A Comparative Study of ChatGPT, DeepSeek, Gemini, and Grok Models

By Dr Wong Wei Kei

University Malaya Medical Center, Malaysia

Introduction

- After a kidney transplant, patient will need lifelong immunosuppressants and this necessitates accessible, accurate, and empathetic patient education.

- Generative artificial intelligence (AI) models have emerged as tools to assist in patient interaction
  - However, its role in patient education on immunosuppression post-kidney transplantation remains uncertain.

Aim

- This study aims to compare the performance of four accessible AI models as stated below in handling questions related to immunosuppression post-kidney transplantation.
    1. ChatGPT (3.5)
    2. Gemini (2.0 Flash)
    3. Grok (3)
    4. DeepSeek (V3)

Methods

- 11 standardized immunosuppression post-kidney transplantation questions were input into each AI model
  - with a response word limit of 300 on 2 occasions, 14 days apart back in March 2025
- The outputs were anonymized and independently evaluated by 3 nephrologists and 1 pharmacist
  - 5-point Likert scale across five domains:

Appropriateness

Clarity and comprehensiveness

Personalization and relevance

Consistency

Human-like empathy

Questions/ Prompts used

- All starts with:
  - “I am a kidney transplant patient. I would like to ask……..”

- All ends with:
  - “Kindly limit the response to 300 words”

- Example:

Question 1. I am a kidney transplant patient. I would like to ask why do I need to take anti-rejection medications after a kidney transplant? Kindly limit the response to 300 words

Other questions include

- Can I ever stop taking my immunosuppressive medications after a kidney transplant?
- What should I do if I miss one dose of the anti-rejection medications?
- Common side effects of tacrolimus?
- Common side effects of mycophenolate mofetil?
- Common side effects of steroids (or prednisolone)?
- Do all kidney transplant recipients need steroids?
- Should I get vaccines after a kidney transplant?
- Can I get pregnant or father a child while taking anti-rejection medications after my kidney transplant?
- How do anti-rejection drugs for kidney transplant affect my risk of cancer?
- How do I know if the anti-rejection medications dosage for kidney transplant is enough?

Methods (2)

- Data were analyzed using Friedman's test for ranked data and post-hoc pairwise comparisons via the Nemenyi test.

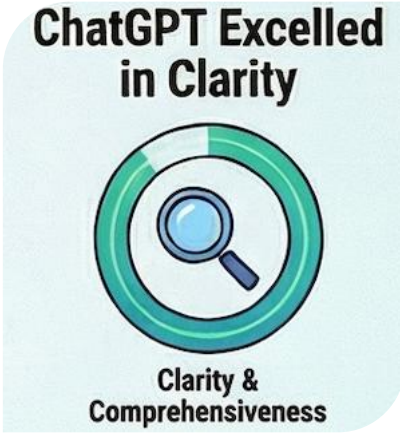
- Inter-rater reliability was assessed using Kendall's W.

Result (1)

Descriptive analysis

- Grok scored highest in most domains
  - Appropriateness 4.48±0.70
  - Personalization and relevance 4.09±0.60
  - Consistency 4.34±0.53
  - Human-like empathy 4.05±0.75

- ChatGPT has the highest mean score in clarity and comprehensiveness domain (4.52±0.51).



Result (2)

- Friedman test demonstrated significant difference between AI models in
  - clarity and comprehensiveness
  - personalization and relevance
  - consistency
  - human-like empathy

Table 1: Friedman test between 4 AI models (ChatGPT, DeepSeek, Gemini, Grok) and post-hoc pairwise comparisons via Nemenyi test

Domains	p-value (Friedman test)	Average ranking (Lower is better)	Post-hoc pairwise comparison via Nemenyi test	q-value (Nemenyi test)	p-value (Nemenyi test)
Appropriateness	0.070	Further analysis not done as Friedman test was not significant			
Clarity and comprehensiveness	**<0.001***	1. Grok (2.15)	DeepSeek - ChatGPT	2.803	0.194
		2. ChatGPT (2.16)	**Gemini - ChatGPT**	**4.262**	**0.014***
		3. DeepSeek (2.70)	Grok - ChatGPT	0.058	1.000
		4. Gemini (2.99)	Gemini - DeepSeek	1.460	0.730
			Grok - DeepSeek	2.861	0.179
			Grok - Gemini	**4.321**	**0.012***
Personalization and relevance	**<0.001***	1. Grok (1.77)	DeepSeek - ChatGPT	2.335	0.349
		2. ChatGPT (2.50)	Gemini - ChatGPT	1.401	0.755
		3. Gemini (2.77)	**Grok - ChatGPT**	**3.737**	**0.041***
		4. DeepSeek (2.95)	Gemini - DeepSeek	0.934	0.912
			Grok - DeepSeek	**6.072**	**<0.001***
			Grok - Gemini	**5.138**	**0.002***
Consistency	**0.014***	1. Grok (2.17)	DeepSeek - ChatGPT	1.343	0.778
		2. ChatGPT (2.47)	Gemini - ChatGPT	0.876	0.926
		3. Gemini (2.64)	Grok - ChatGPT	1.518	0.706
		4. DeepSeek (2.73)	Gemini - DeepSeek	0.467	0.988
			Grok - DeepSeek	2.861	0.179
			Grok - Gemini	2.394	0.327
Human-like empathy	**<0.001***	1. Grok (1.72)	DeepSeek - ChatGPT	0.525	0.982
		2. Gemini (2.70)	Gemini - ChatGPT	0.175	0.999
		3. ChatGPT (2.74)	**Grok - ChatGPT**	**5.255**	**0.001***
		4. DeepSeek (2.84)	Gemini - DeepSeek	0.701	0.960
			Grok - DeepSeek	**5.780**	**<0.001***
			Grok - Gemini	**5.080**	**0.002***
Overall scores	**<0.001***	1. Grok (1.62)	DeepSeek - ChatGPT	2.686	0.228
		2. ChatGPT (2.47)	Gemini - ChatGPT	2.335	0.350
		3. Gemini (2.92)	**Grok - ChatGPT**	**4.321**	**0.012***
		4. DeepSeek (2.99)	Gemini - DeepSeek	0.350	0.995
			Grok - DeepSeek	**7.006**	**<0.001***
			Grok - Gemini	**6.656**	**<0.001***

*Statistically significant ($p<0.05$)

Result (3)

- Post-hoc comparison showed that in Grok is significantly better in:
  - Clarity and comprehensiveness
    - (Grok > Gemini)
  - Personalization and relevance
    - (Grok> Chat GPT, DeepSeek, Gemini)
  - Human-like empathy
    - (Grok> Chat GPT, DeepSeek, Gemini)

Table 1: Friedman test between 4 AI models (ChatGPT, DeepSeek, Gemini, Grok) and post-hoc pairwise comparisons via Nemenyi test

Domains	p-value (Friedman test)	Average ranking (Lower is better)	Post-hoc pairwise comparison via Nemenyi test	q-value (Nemenyi test)	p-value (Nemenyi test)
Appropriateness	0.070	Further analysis not done as Friedman test was not significant			
Clarity and comprehensiveness	**<0.001***	1. Grok (2.15) 2. ChatGPT (2.16) 3. DeepSeek (2.70) 4. Gemini (2.99)	DeepSeek - ChatGPT	2.803	0.194
			Gemini - ChatGPT	**4.262**	**0.014***
			Grok - ChatGPT	0.058	1.000
			Gemini - DeepSeek	1.460	0.730
			Grok - DeepSeek	2.861	0.179
			Grok - Gemini	**4.321**	**0.012***
Personalization and relevance	**<0.001***	1. Grok (1.77) 2. ChatGPT (2.50) 3. Gemini (2.77) 4. DeepSeek (2.95)	DeepSeek - ChatGPT	2.335	0.349
			Gemini - ChatGPT	1.401	0.755
			Grok - ChatGPT	**3.737**	**0.041***
			Gemini - DeepSeek	0.934	0.912
			Grok - DeepSeek	**6.072**	**<0.001***
			Grok - Gemini	**5.138**	**0.002***
Consistency	**0.014***	1. Grok (2.17) 2. ChatGPT (2.47) 3. Gemini (2.64) 4. DeepSeek (2.73)	DeepSeek - ChatGPT	1.343	0.778
			Gemini - ChatGPT	0.876	0.926
			Grok - ChatGPT	1.518	0.706
			Gemini - DeepSeek	0.467	0.988
			Grok - DeepSeek	2.861	0.179
			Grok - Gemini	2.394	0.327
Human-like empathy	**<0.001***	1. Grok (1.72) 2. Gemini (2.70) 3. ChatGPT (2.74) 4. DeepSeek (2.84)	DeepSeek - ChatGPT	0.525	0.982
			Gemini - ChatGPT	0.175	0.999
			Grok - ChatGPT	**5.255**	**0.001***
			Gemini - DeepSeek	0.701	0.960
			Grok - DeepSeek	**5.780**	**<0.001***
			Grok - Gemini	**5.080**	**0.002***
Overall scores	**<0.001***	1. Grok (1.62) 2. ChatGPT (2.47) 3. Gemini (2.92) 4. DeepSeek (2.99)	DeepSeek - ChatGPT	2.686	0.228
			Gemini - ChatGPT	2.335	0.350
			Grok - ChatGPT	**4.321**	**0.012***
			Gemini - DeepSeek	0.350	0.995
			Grok - DeepSeek	**7.006**	**<0.001***
			Grok - Gemini	**6.656**	**<0.001***

*Statistically significant ($p<0.05$)

Result (4)

- Overall performance also differed significantly
- Grok ranking significantly higher than ChatGPT, Gemini and DeepSeek.
- The average ranks of Grok (best), ChatGPT, Gemini and DeepSeek (worst) were 1.62, 2.47, 2.92 and 2.99 respectively.

Table 1: Friedman test between 4 AI models (ChatGPT, DeepSeek, Gemini, C

Domains	p-value (Friedman test)	Average ranking (Lower is better)
Appropriateness	0.070	Further analysis not d
Clarity and comprehensiveness	<0.001*	1. Grok (2.15) 2. ChatGPT (2.16) 3. DeepSeek (2.70) 4. Gemini (2.99)
Personalization and relevance	<0.001*	1. Grok (1.77) 2. ChatGPT (2.50) 3. Gemini (2.77) 4. DeepSeek (2.95)
Consistency	0.014*	1. Grok (2.17) 2. ChatGPT (2.47) 3. Gemini (2.64) 4. DeepSeek (2.73)
Human-like empathy	<0.001*	1. Grok (1.72) 2. Gemini (2.70) 3. ChatGPT (2.74) 4. DeepSeek (2.84)
Overall scores	<0.001*	1. Grok (1.62) 2. ChatGPT (2.47) 3. Gemini (2.92) 4. DeepSeek (2.99)

*Statistically significant (p<0.05)

Result (5)

The lack of inter-rater agreement across domains reflects the subjective nature of evaluating AI content.

Discussion

- Grok demonstrated superior overall performance over ChatGPT, Gemini and DeepSeek in our study
- Only Gemini provided the reference that was used for its responses
- We have used the free version of each AI models as it is more accessible to the public
- Bias is reduced by:
  - Standardization of using free version of AI models
  - Blinded responses were given to assessors

Limitations

- This study does not include other AI models in the market
- Assessor fatigue was also commented due to the long responses that assessors need to go through which might affect the scoring of responses (especially the final few questions)

How are we going forward?

- AI field is progressing at a speed of knot
- Constant reassessments are needed as AI models are improving individually
- There might be huge difference between free services or paid services
  - Further studies are suggested to compare between paid and free services

Conclusion

- This study highlights notable differences in the quality of AI-generated responses to immunosuppression post-kidney transplantation questions.

- Sophisticated AI models can be integrated in nephrology education, provided they are guided by continued human oversight to ensure contextual relevance and personalized content.