Accuracy of Natural Language Processors for Patients Seeking Healthcare Information

Alex Lois, MD, MS; Robert Yates, MD; Megan Ivy, MD; Judy Chen, MD; Saurabh Khandelwal, MD; Andrew Wright, MD

Background

- 77% of patients use the internet for health information, resulting in one billion health-related internet searches per day.
- Natural language processors (NLPs), such as ChatGPT, bridge the gap between human language and computer understanding using large language models to predict responses in a chat conversation.



Figure 1. Natural language processors used in this study.

- NLPs are capable of sifting through massive amounts of information and providing succinct responses to questions.
- It is not known how NLPs distinguish the quality or account for bias in its information sources when responding to questions.
- Despite concerns, NLPs are being incorporated into search engines and influencing search results.
- The performance of NLPs in addressing questions about surgical conditions has not been established.

Hypotheses

- NLPs would be capable of providing relevant and accurate responses to questions about inguinal hernia and its treatment.
- NLP-created education materials would be of lower quality and readability than those produced by national surgical societies.

Acknowledgements

- Thank you to our panel of hernia surgeons: Dr. Nicholas Cetrulo, Dr. Colette Inaba, Dr. Roger Tatum, and Dr. Zoe Parr.

Results

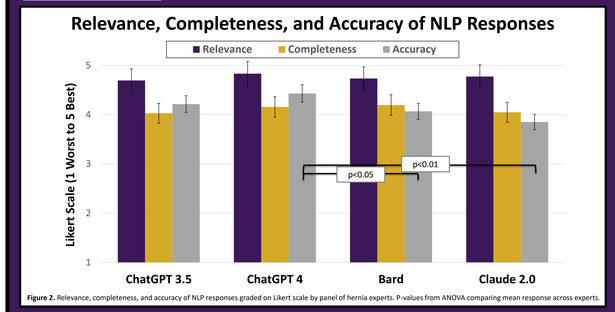


Figure 2. Relevance, completeness, and accuracy of NLP responses graded on Likert scale by panel of hernia experts. P-values from ANOVA comparing mean response across experts.

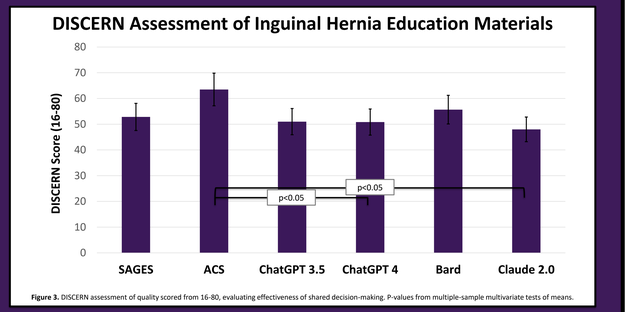


Figure 3. DISCERN assessment of quality scored from 16-80, evaluating effectiveness of shared decision-making. P-values from multiple-sample multivariate tests of means.

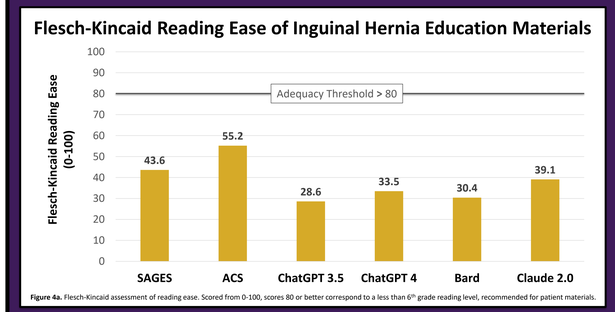


Figure 4a. Flesch-Kincaid assessment of reading ease. Scored from 0-100, scores 80 or better correspond to a less than 6th grade reading level, recommended for patient materials.

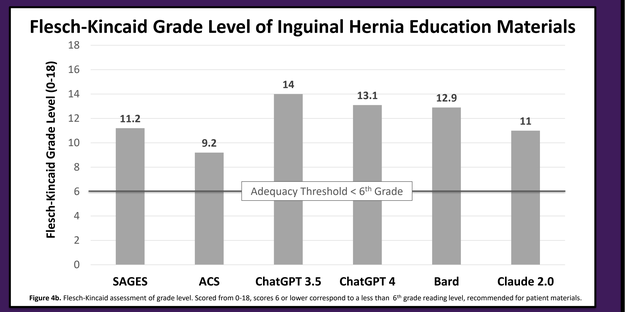


Figure 4b. Flesch-Kincaid assessment of grade level. Scored from 0-18, scores 6 or lower correspond to a less than 6th grade reading level, recommended for patient materials.

Conclusions

- **NLPs provide highly relevant responses of satisfactory accuracy to simulated patient questions about inguinal hernia and its treatment.**
- **NLP-created education materials received lower quality and readability scores than education materials produced by the American College of Surgeons.**
- **NLPs should not supplant the role of physicians in providing patient education but may be a useful tool for educating patients with appropriate supervision.**

Methods

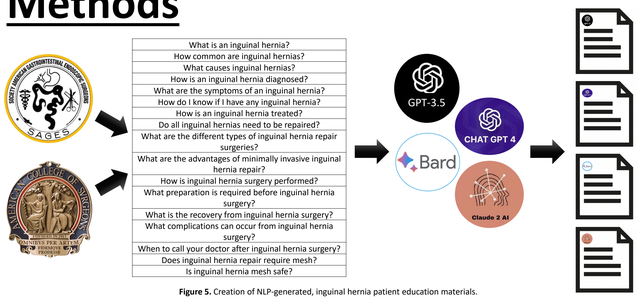


Figure 5. Creation of NLP-generated, inguinal hernia patient education materials.

- Patient education materials produced by SAGES and the American College of Surgeons (ACS) were used as a framework for simulated patient questions.



- NLP responses were collated into education materials and anonymized for evaluation by six hernia surgeons.
- NLP responses to individual questions were scored for relevance, completeness, and accuracy using a Likert scale.
- Compiled NLP education materials were then compared to gold-standard societal materials using two criteria.

FLESCH KINCAID READING EASE

$$206.835 - 1.015\left(\frac{\text{total words}}{\text{total sentences}}\right) - 84.6\left(\frac{\text{total syllables}}{\text{total words}}\right)$$

FLESCH KINCAID GRADE LEVEL

$$0.39\left(\frac{\text{total words}}{\text{total sentences}}\right) + 11.8\left(\frac{\text{total syllables}}{\text{total words}}\right) - 15.59$$

- Readability was assessed using the Flesch-Kincaid reading ease and grade level assessments calculated by Microsoft Word. The AMA recommends a 6th grade level for patient materials.
- Quality was assessed with the DISCERN tool, a validated, 16-question tool for evaluating the effectiveness of consumer health information in assisting with shared decision-making.

Limitations

- Queries were written to simulate patient questions and NLP performance may be improved with more detailed prompts.
- These results represent a single moment in time and as NLPs continue to evolve, performance will likely improve.