Clinical Chemistry

ChatGPT vs Medical Professional: Analyzing Responses to Laboratory Medicine Questions on Social Media

Article:
Mark R Girton, Dina N Greene, Geralyn Messerlian, David F Keren, Min Yu.
ChatGPT vs Medical Professional: Analyzing Responses to Laboratory Medicine Questions on Social Media
Clin Chem 2024; 70(9): 1122–39. https://doi.org/10.1093/clinchem/hvae093

Guest: Dr. Min Yu from Beth Israel Deaconess Medical Center in Boston, Massachusetts.

Bob Barrett: This is a podcast from *Clinical Chemistry*, a production of the Association for Diagnostics & Laboratory Medicine. I'm Bob Barrett.

In recent years, patients have enjoyed increased access to their health information and have become increasingly engaged in their own health care decisions. Most view this as a positive development, but with increased access and engagement inevitably come increased questions. What does this test result mean? What are the symptoms of this condition? Should I choose treatment A or treatment B?

Seeking answers, many patients post health related questions on social media forums, which are then answered by medical professionals. The next question then is, how good are the answers on these social media sites? If the answers provided by humans leave something to be desired, is this an area where ChatGPT and other large language models might have a positive impact?

A new research article, appearing in the September 2024 issue of *Clinical Chemistry*, tackles this issue by comparing the answers provided by human health professionals against those provided by ChatGPT when presented with the same set of health-related questions. In this podcast, we are pleased to speak with the article's senior author.

Dr. Min Yu is the director of Point of Care Testing and Clinical Chemistry at Beth Israel Deaconess Medical Center. She has a professional interest in integrating AI tools into laboratory medicine to enhance clinical practice and patient outcomes. Dr. Yu, just to start off, what motivated you to explore this topic and how did the idea for this study initially come about?

Min Yu: Well, Bob, the idea really comes from noticing a growing trend. More patients are turning to online resources for answers to their medical questions, especially when it comes to understanding their lab results. Well, several factors contribute to this shift.

Well, first, with the rise of standalone labs and direct to consumer testing, patients often receive their lab results

without a healthcare provider to interpret them. And also, the widespread adoption of patient portals now gives people easy access to their lab results, often before speaking to their doctor. And of course, let's not forget the COVID-19 pandemic. It pushed more patients to rely on online resources when in person visits weren't option.

Lab results, however, can be tricky. They're more than just normal or abnormal numbers, and this is why the involvement of a healthcare professional is so important. Well, with more people turning to large language model tools like ChatGPT for answers, we wanted to explore whether these tools could really fill the gap, or if they might actually lead into misunderstandings or even harm if not properly used.

What really caught my attention was a study published in *JAMA* last year that compared ChatGPT's answers to general medical questions posted on social media with those from physicians. From that study, they found that ChatGPT's responses were more preferred and highly rated, particularly for their quality and empathy over those provided by real doctors.

However, that study focused on broad health related questions such as the risks of swallowing a toothpick or potential blindness from bleach exposure rather than the specialized field of lab medicine. This led me to wonder, could ChatGPT handle the complexity and detail required to interpret lab results?

While there have been some studies examining ChatGPT's performance in lab medicine, most have focused on theoretical and controlled setting. We have tested that ChatGPT's ability to interpret simulated reports or answer technical questions, but real-world applications where patient-specific context is crucial remain largely unexplored.

Our study was driven by both curiosity and concern. Could ChatGPT effectively communicate complex lab information in response to actual patient queries? And how would it compare to the responses from human professionals who handle these inquiries every day?

Bob Barrett: Well, that makes lots of sense, especially given how accessible AI tools have become. Could you walk us through how you conducted this study and some of your findings?

Min Yu: Yeah, certainly, Bob. We first gathered real-world patient questions and corresponding responses from medical professionals, and those were thoughts from Reddit and Quora, platforms known for their community driven discussions and expert knowledge sharing.

We focus on a range of topics within lab medicine, including diabetes, thyroid disorders, pregnancy, toxicology, and of course, general chemistry. Then we input those questions into fresh ChatGPT sessions using both versions 3.5 and 4.0, without any tailored prompts. The idea was to replicate how a patient might naturally interact with ChatGPT.

After that, the responses both from medical professionals and ChatGPT were randomized, anonymized, and evaluated by three experienced lab medicine professionals using the REDCap survey platform. And when we analyzed the results, we found that ChatGPT's response were generally longer than those from medical professionals.

Similar to the finding from *JAMA*, we also found those ChatGPT responses were highly rated and preferred by evaluators. The reason for that is because of its accuracy, comprehensive coverage, and relevance. I can give you an example. There's one case where a patient asked about anemia lab results and expressed concern over extreme fatigue. The physician's response was brief, noting that the patient was mildly anemic and should feel better soon with iron supplementation. However, it didn't fully address the patient's symptoms or concerns or in contrast, ChatGPT's response was far more detailed. It not only provided a thorough interpretation of each lab result, but also emphasized the importance of uncovering the underlying cause of the iron deficiency anemia. Additionally, ChatGPT addressed the patient's symptoms with empathy saying, "Oh, I'm sorry to hear that you are feeling this way." This response was appreciated not just for its detailed and accurate explanation of the lab results, but also for its empathy and the professionalism.

This intriguing finding aligns with previous studies, including the one I mentioned earlier published in *JAMA*, which noted that ChatGPT often exhibits a higher degree of empathy compared to the human responses, and this is an area worth exploring further, as it could play a significant role in how large language models integrate into patient care since empathy is crucial in healthcare setting for building trust and ensuring patient satisfaction.

And also, it's important to know that ChatGPT can provide concise answers if prompt, but in our study, we choose not to apply any specific prompts to better simulate real life interactions. Well, another interesting aspect of our findings was that there was no significant difference in preference between ChatGPT versions 3.5 and 4.0. Given that ChatGPT's training does not specifically focus on medical data, the lack of significant improvement between its versions was not surprising to us.

We also noticed some limitations. The medical professional's responses came from a wide range of specialties, not all specific to lab medicine, and this diversity in background likely contributed to the variability in the quality of the responses. However, this also reflects the real-world scenario where patients often receive advice from various medical professionals who may not have specialized expertise in laboratory medicine.

Overall, our study suggests that ChatGPT holds considerable promise in responding to lab medicine related questions, often matching or even surpassing the responses from medical professionals in social media settings. However, there's still a need for more extensive research to validate these findings and to explore how large language model can best be used to support diagnostic process and patient care.

Bob Barrett: It sounds like you're talking about a lot of potential, but some clear limitations too. How do you envision large language models like ChatGPT being used in laboratory medicine?

Min Yu: Let's dive into the specific potential of large language models like ChatGPT in healthcare while keeping a realistic perspective. Starting with the medical education.

Large language models are becoming a valuable resource for creating personalized learning experiences or they can generate a detailed case studies, simulate patient interactions, and provide immediate feedback on medical queries. So, all contributing to more dynamic and effective learning environments.

In clinical practice, large language model has shown promise in reducing clinician workload, particularly in drafting responses to patient inquiries within electronic health record, or EHR, system. For example, a study published this July found that ChatGPT generated drafts were widely adopted by clinicians at Stanford Healthcare. Those clinicians report significant reductions in workload and burnout as a result.

Turning to lab medicine, the applications are a bit more specialized. One area where ChatGPT could really make a significant impact is helping patients understand their lab results through patient portals. As our study found, ChatGPT can generate responses that are not only accurate, but also empathetic to creative qualities that in patient communication is important. However, the key will be ensuring that this model generated explanations are both accurate and easy to understand to our patients.

Looking ahead, large language models' potential in lab medicine could also extend to document preparation, such as

standardizing the writing and updating of SOP, or standard operating procedures. Additionally, it could assist in preparation for regulatory such as CAP inspections by generating reports, ensuring all documentation is correctly formatted and up to date, so thereby streamlining the preparation process.

In short, while there's a tremendous potential, their success will depend on thoughtful integration, continuous evaluation and collaboration with the healthcare professionals.

Bob Barrett: Well, it seems like there's a clear vision for the future, but it raises questions about where research should go next. Looking ahead, what do you think should be the focus of future studies in this area?

Min Yu: Great question, Bob. Future research in this area should focus on a few key priorities. First and foremost, we need to improve the contextual understanding of large language models, which means training models like ChatGPT specifically on lab medicine data. Developing and validating domain-specific large language models, or small language models trained on authoritative medical data, can address current limitations, and this specialized training would significantly improve the accuracy and reliability of information those models generate.

Another important research direction is, of course, rigorous validation of model generated medical information. While current research showed that large language models can offer accurate and empathetic responses, these findings need to be confirmed across diverse patient populations, question types, and clinical settings, and it's especially important to assess how well these models handle complex cases requiring detailed interpretation of lab results.

Finally, addressing the ethical privacy and legal aspects of large language model in healthcare is paramount, ensuring that outpatient data is protected, maintaining transparency in large language model generated recommendations, and establishing clear guidelines for large language model use are essential to ensuring the safety of both patients and practitioners.

Bob Barrett: Well, it is clear there's a lot more work to be done, but also a lot of promise. Finally, Dr. Yu, what do you hope listeners take away from your study and from this conversation?

Min Yu: As we wrap up this conversation, what I hope listeners take away is that while large language models, particularly models like ChatGPT, hold significant potential in healthcare, including lab medicine, they are not a replacement for human expertise. They are a tool to augment it.

Our findings suggest that ChatGPT can provide accurate, comprehensive, and empathetic responses, sometimes even surpassing those of medical professionals in certain scenarios.

However, it's crucial to remember that in their current stage, large language models lack the clinical judgment and the contextual understanding that come with years of medical training and experience.

As we move forward, the integration of large language models in healthcare should be done thoughtfully, with a focus on complementing rather than replacing human touch.

I also hope this conversation highlights the importance of ongoing research and the collaboration between the tech and medical communities. By working together, we can refine those tools, address their limitations, and ultimately improve patient care in ways we are just beginning to explore.

Bob Barrett: That was Dr. Min Yu from Beth Israel Deaconess Medical Center in Boston, Massachusetts. She wrote a research article evaluating ChatGPT's ability to answer laboratory medicine questions in the September 2024 issue of *Clinical Chemistry*, and she's been our guest in this podcast on that topic. I'm Bob Barrett. Thanks for listening.