Clinical Chemistry

Assessing the Accuracy and Clinical Utility of ChatGPT in Laboratory Medicine

Article:
Carlos Munoz-Zuluaga, Zhen Zhao, Fei Wang, Matthew B Greenblatt, and He S Yang.
Assessing the Accuracy and Clinical Utility of ChatGPT in Laboratory Medicine.
Clin Chem 2023; 69(8): 939-40. https://doi.org/10.1093/clinchem/hvad058

Guest: Dr. He Sarina Yang is an Associate Professor in the Department of Pathology and Laboratory Medicine and the Medical Director of Clinical Chemistry and Toxicology Service at New York Presbyterian Hospital/Weill Cornell Medicine.

Bob Barrett: This is a podcast from the journal *Clinical Chemistry*, a production of the Association for Diagnostics and Laboratory Medicine. I'm Bob Barrett. Do you trust your physician to provide correct answers to your medical questions? How about an artificial intelligence program with the ability to generate remarkably humanlike answers? Would you be able to tell the difference between an answer from a human with medical training and one from a natural language processing tool trained on massive datasets? The Chat Generative Pretrained Transformer, or ChatGPT, has provoked considerable interest since its release in late 2022, and can be used to answer questions in any number of areas including disease diagnosis and the interpretation of clinical laboratory test results. While the answers appear convincing, sometimes they're correct and sometimes they're not.

The scary part is that unless the user has experience in laboratory medicine, it's difficult to tell the difference. How well can ChatGPT answer questions related to clinical laboratory testing? Can the power of ChatGPT be harnessed to make a positive impact on patient care or does the potential for harm outweigh the possible benefits? A letter to the editor appearing in the August 2023 issue of *Clinical Chemistry* tackles these questions and today we're excited to talk with the letter's senior author. Dr. Sarina Yang is an associate professor in the Department of Pathology and Laboratory Medicine and the Medical Director of Clinical Chemistry and Toxicology Service at New York Presbyterian Hospital/Weill Cornell Medicine. She is also a member of the IFCC Working Group on Artificial Intelligence and Genomic Diagnostics. So, Dr. Yang, could you tell me about your recent publication on ChatGPT and lab medicine?

Sarina Yang: Yes, definitely. ChatGPT is a chatbot service developed by OpenAI. It is a large language model driven by artificial intelligence and trained on massive amounts of data. As you said, ChatGPT has gained tremendous attention since its release in November 2022 because it has shown an extraordinary ability to generate humanlike answers in response to text input in a conversational context. While there is a widespread interest in the use of ChatGPT, its

application in medicine and health care is still controversial. ChatGPT was not extensively trained on biomedicine data, nor were its responses thoroughly assessed by medical experts. However, patients and clinicians may turn to ChatGPT for assistance in interpreting laboratory data or understanding how to utilize clinical lab services.

Therefore, we carried out a pilot study to evaluate ChatGPT's ability to answer questions frequently encountered in the lab medicine field, ranging from basic medical or technical knowledge, interpretation of laboratory data in clinical contest, FDA regulation, to laboratory operation. We asked 65 questions in ChatGPT including clinical chemistry, toxicology, transfusion medicine, microbiology, molecular pathology, hematology, coagulation, and laboratory management. The sources of questions included textbooks, clinical pathology board questions, clinical consultation questions, and real-life case scenarios. Then, three faculty members evaluated ChatGPT's responses and scored them as either fully correct, or partially correct, or totally incorrect, or irrelevant.

We first performed the evaluation using the ChatGPT-3.5 version and found ChatGPT correctly answered 26% of the questions, generated an incomplete or partially correct answer in 31% of the questions, was totally incorrect in 34% of the questions, and gave an irrelevant answer to 9% of the questions. For example, ChatGPT cannot differentiate respiratory versus metabolic acidosis when given a panel of blood test results, and it cannot explain why urine toxicology results maybe falsely negative when urine creatinine level is less than 10 milligram per deciliter. Sometimes the answer appeared convincing through the use of appropriate terminology but when we looked closely into the answers, many contain inaccuracies. Often, ChatGPT's response was too broad and not sufficiently tailored to the cases or clinical questions to be useful for clinical consultation.

In March, a newer GPT-4 version was rolled out so we re-evaluated all the questions using GPT-4. Well, GPT-4 has indeed demonstrated a notable improvement in accuracy compared to the GPT-3.5 version. GPT-4 correctly answered around 51% of the questions, generated an incomplete or partially correct answer in 23% of the questions, was totally incorrect or misleading in 17% of the questions, and gave an irrelevant answer to 9% of the questions. It performed better with the questions related to basic medical knowledge but less accurate in interpreting complex lab results or explaining FDA regulations. For example, when we asked if there are FDA approved high-sensitivity troponin point-of-care device on the United States market, it said yes and gave a few seemingly convincing examples but those were

hallucinations. These data and examples were shown in our new publication on ChatGPT.

Bob Barrett: Okay, so based on your study, can physicians and nurses ask ChatGPT if they have questions about the use of laboratory tests or if they need help interpreting test results?

Sarina Yang: Overall, I think that ChatGPT is a very powerful tool that has the potential to provide faster responses to routine clinical lab questions. However, some of its answers may not be accurate or tailored to specific clinical consultation questions. For example, it was not able to distinguish alcoholic hepatitis from other types of hepatitis when given a panel of liver enzyme results, resulting in a lengthy response that was not tailored to the case. Also, it may not be able to provide information related to a specific hospital setting. Instruments and methodology dependent information such as the cross-reactivity profile of urine drug immunoassays may be crucial for interpreting certain laboratory test results.

However, such data has not been full fully integrated into ChatGPT's training regimen. Therefore, ChatGPT should not be relied upon without expert human supervision. I think physicians and nurses should still consult the laboratory regarding the use of laboratory test, the specific requirement of sample preparation, and interpretation of lab test results. So, clinicians should be vigilant of the risk of using ChatGPT and be very careful with ChatGPT's answers.

Bob Barrett: Is ChatGPT a useful tool for pathology residents when they prepare for their board exam?

Sarina Yang: Well, first, ChatGPT has successfully passed the United States Medical Licensing Exam so it demonstrates the potential to assist with medical education. From our pilot study, it correctly answered approximately 60% of questions related to basic medical or technical knowledge. Honestly, I think some of the answers were pretty good. However, some of the responses were incomplete or inaccurate so it really requires a deep understanding of lab medicine to distinguish between correct and incorrect elements of its responses. In addition, both GPT-3.5 and GPT-4 were trained on data through September 2021 so their answers may not be up to date. For example, ChatGPT does not know the most recent reference value of blood lead in children, which should be 3.5 milligram per deciliter, because this value was updated by the CDC in October 2021.

Furthermore, one thing to point out is that the quality of ChatGPT's response depends on the quality of prompts used. Sometimes, rephrasing the query or asking a more specific question leads to a different answer. Also, ChatGPT may generate different responses to the same question when used

by different users. This variability in responses is undesirable for use as a medical reference where consistency is important. So I wouldn't recommend pathology residents to totally rely on ChatGPT when preparing for their CP board exam but it could be a useful tool to help residents remember things that they have already studied.

Bob Barrett: Dr. Yang, what is your vision regarding the application of AI chatbots in lab medicine?

Sarina Yang: In my opinion, AI chatbots holds great potential to improve medical education, provide timely responses to clinical inquiries regarding laboratory tests, assist in interpreting laboratory results, and facilitate communication among patients, physicians, and laboratory professionals. However, their application in lab medicine remained limited so far due to the limitations and challenges that I just talked about. A chatbot suitable for clinical practice should be reliable, manageable, and comprehensible. Nevertheless, current chatbots have not reached a satisfactory level of clinical performance and consistency.

In addition, the current version of ChatGPT may not be able to deliver accurate information on specific areas of clinical specialty or disciplines where advanced knowledge, reasoning ability, and in-depth comprehensive analysis of patient information are required. ChatGPT still lacks the cognitive capability for thinking and reasoning that human possess. Consequently, it is unable to perform novel, logical, and explainable reasoning when encountering previously unseen situations. This limitation affects its capability to generate accurate response to complex queries such as clinical consultation questions.

Bob Barrett: Well, finally, what's next? Where do you think research in this field is headed?

Sarina Yang: Well, I anticipate future development will involve hybrid approaches to address these limitations. Combining large language models with domain-specific modules that are trained on vigorously validated medical resources such as articles, reviews, clinical guidelines, and databases. ChatGPT's conversation should be grounded on reliable data source and can be verified against relevant existing knowledge before being presented to users. To be effectively used in medicine, ChatGPT must be thoroughly evaluated against a standard clinical practice.

Also, in the era of personalized medicine, many inquiries from treating physicians concerning laboratory tests are based on a specific patient's medical presentation rather than the general information listed in the test directory. AI chatbot that can integrate a patient's medical history and laboratory

test results to provide personalized data interpretation would be highly valuable to physicians.

Bob Barrett: That was Dr. Sarina Yang from New York Presbyterian Hospital and Weill Cornell Medicine. She wrote a letter to the editor on the utility of ChatGPT and laboratory medicine in the August 2023 issue of *Clinical Chemistry*, and she has been our guest in this podcast on that topic. I'm Bob Barrett. Thanks for listening.