Clinical Chemistry

Article:
Nicholas C Spies.
Embracing Generative Artificial Intelligence as a Support Tool for Clinical Decision-Making.
Clin Chem 2025; 71(11): 1178–9. https://doi.org/10.1093/clinchem/hvaf084

Guest: Dr. Nick Spies from the University of Utah and the Applied Artificial Intelligence group in the Institute for Research and Innovation at ARUP laboratories in Salt Lake City, Utah.

Bob Barrett: This is a podcast from *Clinical Chemistry*, a production of the Association for Diagnostics & Laboratory Medicine. I'm Bob Barrett. Generative artificial intelligence tools have revolutionized many aspects of daily life with comments like "I'll just ask ChatGPT" or "ChatGPT can do it for me" becoming commonplace. Predictably, interest is spilled over into the medical world, with several studies comparing generative AI against humans with medical training asking whether humans or AI do a better job of answering a series of clinical questions. While many studies have pitted the two head to head, this may not be the right question to ask as it's unlikely that we'll replace human physicians with AI tools anytime soon. Perhaps the better question is whether AI can make human physicians better.

Essentially, do physicians who use AI perform better than physicians who only access conventional resources? A News & Views article, appearing in the November 2025 issue of *Clinical Chemistry*, explains a research study that asked exactly this question and explores findings that could translate to the clinical laboratory. In this podcast, we're joined by the article's lead author. Dr. Nick Spies is an Assistant Professor of Pathology at the University of Utah and medical director of the Applied Artificial Intelligence Group in the Institute for Research and Innovation at ARUP Laboratories. So, Dr. Spies, we seem to be hearing more and more about generative AI and large language models in medicine. Could you provide a quick background on what these are and how they work?

Nick Spies: Yeah, of course. So, generative AI is so called because instead of our classical artificial intelligence approaches, which just provide kind of one number of an output or a classification, generative applications can generate entire either sentences or paragraphs of text. They can produce images. They really have kind of no cap on the length of the outputs that they produce. And so, generative AI itself is kind of a pretty big umbrella under which, I think, maybe the most common use case so far has been the large language model, which obviously takes in language and then produces language.

Under the hood, these large language models are really just interpreting the language inputs that they're taking in and converting them to numbers, which they use as kind of a similarity-based metric to produce output that would fit the expectations for human language based on whatever inputs they have. And there's been a lot of kind of fancy applications and add-ons to these models that have really fine-tuned them to sound and feel very much like human speech or human conversation. But it is important to note, however, that these language models, they're fine-tuned to produce human sounding text, but not necessarily factually correct texts.

So, a lot of these applications, they look very much like correct appropriate responses on the surface. But when you dig a little bit deeper, you can start to see some of these cracks in the surface, if you will. And it's really important that when we're using these models for medical applications or for laboratory applications that we are very robust to the fact that factuality is something that we'll have to take a really keen note of.

Bob Barrett: And how have these tools been applied to laboratory medicine so far?

Nick Spies: Yeah, I think we're very much in the exploratory phase of these technologies as far as what they can kind of do for us both in our personal and professional lives. But professionally at least, there have been a lot of interesting papers published so far asking kind of very simple question and answer type tasks -- you know, where we'll give some examples of laboratory results and we'll ask the LLM to interpret these results. And then at least in the published studies -- you know, we've had human graders that have evaluated the LLMs responses based on factuality and accuracy and completeness and -- you know, ability for interpretation errors to cause danger. And largely that is most of what the published literature has looked like so far.

That said, there have been some really useful and interesting works for information extraction or information retrieval type tasks. We go to conferences and we hear a lot of people talking about how pathology reports being largely unstructured text have been very difficult to do big -- large scale research projects on, until we have these large language models that are coming out that allow us to kind of extract diagnoses and disease burdens and all sorts of other things from these big chunks of text, so that we can do kind of more structured research at scale. And so in a big way, I think the act of retrieving more structured, easy to use inputs out of largely unstructured text has been the main use case for LLMs in the laboratory so far.

Bob Barrett: All right. Well, that's a good setup to discuss the News & Views article. Could you fill us in on the primary article you discussed and what did they set out to study? How did they do it?

Nick Spies: Yeah, absolutely. So, the obvious next step beyond these relatively straightforward information retrieval tasks would be to ask the question, can these large language models provide appropriate clinical reasoning and clinical reasoning in a broad scale -- you know, for our medicine colleagues means answering diagnostic vignettes. And so, in this really interesting article from the Jonathan Chen lab out in Stanford, they basically did a randomized controlled trial approach using ChatGPT and then kind of standard of care reference texts -- you know, like, Harrison's Internal Medicine textbook [Harrison's Principles of Internal Medicine] or for us the Tietz Clinical Chemistry [Tietz Fundamentals of Clinical Chemistry and Molecular Diagnostics] textbook.

And they compared the two and then they asked ChatGPT alone those identical kind of clinical vignettes. And so, we have this three pronged approach where we have these bespoke kind of real world use cases that the authors develop themselves so that we can feel pretty confident that they aren't in the training data sets for these large language models. And we basically just ask sequential diagnostic decisions. You know, a 78-year-old male presents with a cough. He was a 30 pack year smoking history. What should we do next? Multiple choice. There's a correct answer and several incorrect answers.

And what they found actually was quite striking in that the worst of their three performance groups was in the physician set that had no generative AI help to it. All they had was the reference standard of care materials. Above that group, they had the physicians that could use ChatGPT to help them answer the questions. And then actually even more strikingly, above the physicians with AI help, the ChatGPT alone arm that only had the large language models answering the questions, actually had more consistently correct answers than the physician groups themselves.

Bob Barrett: This is all very interesting. How would you interpret these results more broadly?

Nick Spies: Yeah, I think there's a lot to dig into here for sure. I think the biggest takeaway point here is that while it is very simple, very convenient, a very nice conclusion to make if we do say something along the lines of AI is great, humans are great, but humans and AI are probably best. There's more and more published literature coming out to suggest that that's maybe not always the case. Even in the laboratory itself, we have kind of examples of machine learning approaches.

A good one is from Dr. Farrell down in New Zealand that showed that when we're looking at wrong-blood-in-tube errors, if we add a human to the machine learning task, the performance actually does decrease a little bit. And that's certainly not to say, especially in the case of clinical decision making that we should be replacing humans entirely with these language models. But it does kind of lean into this idea that from a performance perspective alone, these large language models are getting really, really good at what they do in these pristine, sandbox, perfect use case type examples like this question and answering clinical trial that we're reporting on here.

Bob Barrett: Is there anything that laboratorians can take from this that might transfer to their own clinical work?

Nick Spies: Yeah, absolutely. I think we should, as laboratorians, get more and more comfortable with what these models can do and what they know and what they don't know, obviously. But from this first results comparison in that the physicians with the standard of care reference textbooks underperformed the physicians with ChatGPT access, it's very likely that there will be plenty of clinical dilemmas within the laboratory that if we are familiar and facile with the LLM technologies that are available to us, we can probably improve how we manage some of these tasks in our own day to day lives. That's not to say that we should be fully automating any of our complex decision making, especially when it comes to really kind of complex ethical and legal arguments or considerations.

You know, the classic example comes up of if a mistake is made, who should be liable? I don't think any of us are going to be having any success suing the Googles or the OpenAI's of the world if they help us make a bad decision. And so, the ultimate responsibility will still obviously land on the medical directors or the laboratorians to use all of the information at their disposal to make correct decisions. But it will certainly be more and more likely that these really powerful tools can help us do our day to day tasks or even these really powerful tools can automate some of the lower risk, more rote, low complexity tasks, if you will.

And so as laboratorians, we should really start to kind of invest the time and energy to get very familiar with these technologies and start to kind of build out our own sandboxes, if you will, to see what of these technologies can we leverage in our day to day work to improve what we do both from an operational efficiency standpoint and from a patient care clinical quality standpoint.

Bob Barrett: Well, finally, Dr. Spies, let's look ahead. Where do you think the field should go from here?

Nick Spies: Yeah, so I think the biggest place to go, you know, in terms of where this paper is really giving us a good guide, is building out these robust clinical evaluation frameworks for our own tasks. So, when we mention things like "how are these models going to be helpful for us in the laboratory?" What we don't want to do is just open up the public-facing Internets and type in our patient's information and take the first thing that ChatGPT spits out as gospel and starting to make clinical decisions based on it.

What we really need is to build out these really nice, published, peer reviewed examples of cases that have been really tightly vetted, both from a technical standpoint where we're making sure that our information security is correct and how we're handling the data is appropriate. And from a clinical quality standpoint where we're saying, okay, we have a highly reproducible use case, we know what the inputs look like and we have a verifiable output, which means whenever the large language model produces some kind of output, we can check how correct it is, or if it is correct. And those are the kinds of tasks that this paper gives us a great guide into asking in that the paper is really structured largely like a clinical trial.

We have three comparison groups. You know, we have the physicians alone or the physician's standard of care. We have physicians plus ChatGPT, and we have ChatGPT alone. That kind of framework where we have our current state or our standard of care and then our standard of care plus AI interwoven in what we would consider a human-in-the-loop type framework, or the fully automated LLM alone strategy, will be a really good way to compare how we can get the most value out of these models without subjecting our patients to undue risks or our patient data to unnecessary security leaks.

Bob Barrett: That was Dr. Nick Spies from ARUP in Salt Lake City, Utah. He wrote a News & Views article in the November 2025 issue of *Clinical Chemistry,* describing potential benefits of using generative artificial intelligence as a tool to guide clinical decision making, and he's been our guest in this podcast on that topic. I'm Bob Barrett. Thanks for listening.