AI-GR Pod 23 10.11.24 David Ouyang

[00:00:00] Half the challenges of running RCT aren't entirely technical. In fact, I would say 80% revolve around the logistics as well as the human aspects. So that involves kind of the IRB approval. We had to work with the Cedars-Sinai IRB so that they were comfortable with the amount of consent we're providing.

We thought it was minimal risk because we were not impacting the sonographer in any way. And also the cardiologists finalize everything so that it's ultimately the cardiologist's call. The technical implementation was important in that we directly integrated with the software that the clinicians use to assess echo.

This is called PAC system or structural reporting system where we wanted to show the AI model tracings in the exact same way that sonographers and cardiologists see it. So, Brian Hie, one of the Stanford Ph.D. students, actually directly embedded the model into syngo Dynamics, our PAC system. But the overall highlight, and I think that this was a Twitter post, which [00:01:00] people asked me, how much did this cost?

This study cost roughly 20 bottles of wine, 50 Starbucks gift cards, and then my effort. That's amazing.

Welcome to another episode of *NEJM AI Grand Rounds*. I'm Raj Manrai and I'm excited today to bring you our conversation with Dr. David Ouyang. David is a cardiologist and researcher at Cedars-Sinai Medical Center in Los Angeles, and he's been at the forefront of applying AI to cardiology for many years.

He took us behind the scenes of some of his biggest papers in this space, and it was really eye opening to hear how creative and clever he was in launching a clinical trial at such an early phase of his career. This is a career stage where I think it's tempting to say that running a clinical trial is something I can only do later, when I'm much older, but he really countered that narrative, and he showed us just how he did it.

He also told us about his foundation in statistics and his time working with some of the giants in the field. And I think this really [00:02:00] shines through now in his approach to AI and medicine.

The *NEJM AI Grand Rounds* podcast is brought to you by Microsoft, Viz.ai, Lyric, and Elevance Health. We thank them for their support.

And with that, we bring you our conversation with Dr. David Ouyang. Well, David, we're super excited to have you with us today on *AI Grand Rounds*. Thanks for coming. Yeah. And Raj, thanks so much for the invitation. I'm really glad to be here. David, this is a question we ask all of our guests right at the beginning.

Could you please tell us about the training procedure for your own neural network? How did you get interested in AI and what data and experiences led you to where you are today? So, I was born in China, grew up in Texas, spent some time in Houston, Dallas, as well as Austin. Went to Rice University, majored in statistics and biochemistry.

My statistics advisor was actually [00:03:00] Hadley Wickham, who was the creator of ggplot and tidyverse. So really learned a lot from him. Went to UCSF for medical school, initially thought I wanted to do an M.D., Ph.D., but ultimately decided on doing M.D. And then was at Stanford for residency and fellowship.

That's where I met a lot of close collaborators and mentors, people you guys already had on the podcast, including Ewan and James. And then, now I'm currently an assistant professor at Cedars-Sinai Medical Center. So, we always like to pull it back a little bit further than that.

So, what, intrigued the young David about statistics and medicine and how did you kind of like get set on that path in the first place? Let's go beyond the CV bio for just a sec. So, I'll do a brief tangent, but have you guys seen the movie *Arrival*? Yeah, the high-level overview is that aliens come to earth.

But they teach the humans a language, and that completely changes their perspective, right? In that movie, the language doesn't have a beginning or [00:04:00] end, and that they can actually then see into the future in the past. But I really believe that kind of how we think about questions, whether in science or how we apply them, is really based on our training.

Obviously, medicine training is a crucible. It's one of those things where, by sheer nature, being the only person there at the end of the night, you learn a lot of medicine. But I would say that I've really had a lot of value and a lot of my perspective comes from working with Hadley in terms of data visualization, knowing how, I think variables are structured, how you actually want to present

it, how you can potentially hide things, how you want to elevate things actually has a, gives me a lot of perspective on what problems I choose for research and what are tractable AI problems.

I can't believe I didn't know that you had worked with Hadley Wickham. That's such an amazing factoid. I remember his ascendancy in the early 2010s when I was a grad student, and he had the grammar of graphics and I always thought he was like the coolest statistician, that I had ever heard of.

And [00:05:00] so now like in hindsight, it makes complete sense that of course you worked with him because I think you're also like one of these very cool nontraditional statisticians. So, it's like, *Arrival*, it's like all coming together for me now. Fun fact, and I think Hadley should speak for himself, but Hadley actually went to medical school.

He is from New Zealand, went to medical school and ended up deciding medicine's not for him. And he was a professor at Rice. And I think similar falling through your footsteps, Andy, I think ended up going to our studio. I think they now have a new name and went to really focus on leveraging kind of technology for data science.

And maybe just some context here for our medical listeners who don't know who Hadley Wickham is. So, Hadley had, has launched a revolution in data science. He has created a suite of R packages for data visualization but has also reinvented how we visualize data in the first place. He's so influential that like one of my good friends actually named his daughter Hadley, in recognition of Hadley Wickham's contribution.

So, a complete legend in the field of data science. Yeah, when I was in [00:06:00] medical school, he would go to all these tech companies to give master classes. And he would actually need a research assistant or someone to help walk through the crowd. And so, I was really fortunate to, actually, we went down to eBay and I helped some of their full time people learn, ggplot and stuff like that.

So that was actually a really fun experience and also part of my initial interest in technology and I guess tech companies. David, that's amazing. Like Andy, I can't believe I didn't know that you sort of had this, uh, this arc and this, this, uh, kind of research experience working with Hadley. Could you tell us, my guess is that it might be R, but it might be something, uh, totally different and preceding sort of your, your interaction with, with him.

But what was your introduction to programming and data science and computer science? Did you start programming in high school or was this with Hadley in college? Yeah, so fun fact. I went to a math and science high school. I went to the same high school as Roxana. That's where I learned C++, Java. It was only during undergrad [00:07:00] where it was primarily R and Python. And obviously with deep learning in the last couple of years have been much more Python specific.

I would still say that for R manuscripts that I handle the figures for, I still primarily do it in ggplot. So, another like unknown connection to me that you and Roxana, another guest of the podcast and editor at *NEJM AI*, went to high school together. That's amazing. Amazing. So, David, we want to dive now, deep into some of your research.

So, you're a physician scientist, you're a practicing cardiologist, a deep learning researcher, and an echocardiographer. And I think Andy would agree with me that you're among probably the very best examples of anyone that either of us know to truly blend the clinician and researcher parts of your life, the sort of physician scientists, both sides there.

You know, you do many things, but I think one of the recurring topics in your papers is about applying AI to cardiology and in particular, imaging studies and echocardiograms. I think the very [00:08:00] first paper of yours that I read was a paper that you published in *Nature* back in 2020. And the title of the paper is "Video-based AI for beat-to-beat assessment of cardiac function."

So maybe we could start with that one. Could you tell us about what that paper is all about and maybe the backstory as well. How that project got started, where you were, and if you could frame also where the field was at at the time and what you were trying to do with that particular paper.

Yeah, Raj, thanks so much for the comments. You're, you're very kind and it's a very generous introduction to the paper.

When I was a cardiology fellow, I realized that a big part of cardiology practice is in interpreting images and videos. Echocardiography is the most common form of cardiac imaging, and it's something that overnight I had to be called to do and to interpret, oftentimes with backup, but many times feeling more and more comfortable that I could do the assessment.

When I was doing this as a cardiology fellow, I realized that we actually [00:09:00] have large imaging databases that has tremendous information that

spans essentially, the full spectrum of disease, from healthy individuals getting echoes, to patients with very sick hearts in the ICU, as well as needing heart transplant.

And this was back in the era, I would say maybe from 2018 to 2020, when convolutional neural networks were getting much more excitement. The caveat is that at that time, there were very few video-based architectures. Most CNNs are image-based as opposed to video-based. So that temporal relationship and information was not well captured.

And we thought that echocardiography, or cardiac ultrasound, is very inherently a video-based modality, and it's really the place to actually do medical AI that integrates both very technical innovations, as well as applications that leverages those technical innovation. So cardiac function, or left ventricular ejection fraction, is the most important and [00:10:00] standard assessment in echo.

It's how we decide whether someone has heart failure or not. It's how we decide whether someone can still get chemotherapy or not. If there's toxicity and a variety of many other reasons. And that's something that's a very both visual as well as temporal assessment. So, our paper in 2020 was using kind of 3D ResNets, kind of R2 plus 1D, R3D, as well as other approaches to assess heart function, and assess this at many places. And that's where we already did external validation at Cedars-Sinai to show that this is a generalizable model, but also showed that if you're able to use a video-based model, hopefully you can actually have more precision than the cardiology assessment of these measures.

David, for the ML folks who are listening to this, can you just quickly explain what ejection fraction is? Yeah, the left ventricular ejection fraction is a ratio of the heart size when it is full of blood and when it pumps all the blood out. That's our crudest and most common [00:11:00] assessment of heart function, meaning that a really strong heart squeezes roughly about half the blood out in any given heartbeat.

And a really weak heart squeezes maybe 10% of the blood and potentially need multiple heartbeats to have the same amount of blood going through the rest of your body. It's a quantitative measurement. So, it's a number between 0-100% but it's also a very video-based measurement where you need to capture that relationship where the ratio of the largest and smallest sizes to assess that function.

Cool. David, that's amazing. So, you know, in that paper, right, you're applying this neural net architecture to videos. And I think, you know, you hit it there. That's the sort of key technical methodological contribution of that paper.

And then you're setting up a task that is how clinicians themselves, right, how echocardiographers, sonographers interpret the studies of videos, not single images like the sort of previous era of AI applied to similar tasks. Could you tell us maybe about the clinical significance of some of the [00:12:00] findings there?

Like, was it a sort of very clear improvement? Was it obvious to you that this was working right out of the box? Or did you have to sort of refine the task iteratively and even composing that original *Nature* paper to know that this was working or to demonstrate that this had potential?

Yeah, we developed both regression as well as segmentation models, meaning that we had a model that both identified the left ventricle, so it can actually trace the part of the heart that is relevant for this measure, as well as come up with a number that actually relates the assessment of heart function.

When I first saw the model output for segmentation, I was like, wow, that is actually something that is both incredibly laborious that sonographers and cardiologists have to do. And it's not clear that I can tell the difference between me doing it versus the AI model doing it. There's a bunch of reasons why it's better when software or AI does it.

It can do it every single frame and every single beat. It can be more precise and more reproducible. It never gets [00:13:00] tired. It never gets tired. And it ultimately, I think we, we show in the paper that it actually allows for a more precise assessment. And so where is that work now? So, you know, fast forward a few years, are you applying that same algorithm in your hospital and other hospitals that you work with?

Or has it sort of, was it an interesting kind of research study that led to then the next incarnation of these models to be applied in practice? Yeah, so the next step in our assessment after the paper was published, one, we made sure that all the code and all the underlying training data was publicly released.

So as part of the Stanford AMI datasets, we were really fortunate that we've had many other people use it, but we also wanted to see what it looks like in clinicians hands. So, in 2022, we conducted a blinded RCT of this assessment. This is when I already moved to Cedars-Sinai. So, we asked the Cedars-Sinai

echo lab for the sonographers to trace about 4,000 [00:14:00] echoes, showed the cardiologist 2,000 of those, where the other 2,000 we swapped out what was being assessed by AI and conducted a blinded pro-selective RCT.

That was when we showed that clinicians, one, can't tell the difference between AI and sonographer, and two, are less likely to change that assessment. So, it was, while it was a non—. When the AI had generated it. Yes, we were looking at the difference in assessment from the preliminary assessment from the sonographer or AI and the final cardiologist assessment when you ask them to change it to what they think is the ground truth.

So, I'm jumping ahead because we want to, you know, this is really unique. I think you're a machine learning researcher and AI researcher who's been able to lead a RCT where we have so many questions about how you do that, how you get that off the ground. But maybe while we're talking about this particular study, can you just tell us about, what was the, what were the sort of hurdles to even launch this kind of a study at Cedars-Sinai?

I guess maybe one way to frame it is, you know, Andy and I like [00:15:00] to talk about the socio-technical challenges of doing research in academic medical centers and universities. And sometimes there's technical problems, right, technical challenges that you have to solve, and oftentimes there's sort of social ones, where whether it's organizing people around a particular task, getting approval for a study, getting funding, assigning that sort of funding to a particular study. How did you navigate those challenges and what were sort of the main thing that you solved to even launch that study?

Let alone then conduct it and publish it. Yeah, Raj, this is a really great question. Half the challenges of running an RCT aren't entirely technical. In fact, I would say 80% revolve around the logistics as well as the human aspects. So that involves the IRB approval. We had to work with the Cedar-Sinai IRB so that they were comfortable with the amount of consent we're providing.

We thought it was minimal risk because we were not impacting the sonographer in any way, and also the cardiologists finalized everything so that it's ultimately the cardiologist's call. The [00:16:00] technical implementation was important in that we directly integrated with the software that the clinicians use to assess echo.

This is called PACS system, or Structural Reporting System, where we wanted to show it. The AI model tracings in the exact same way that sonographers and cardiologists see it. So, Brian Hie, one of the Stanford Ph.D. students, actually

directly embedded the model into syngo Dynamics, our PAC system. But Raj, I think the overall highlight, and I think that this is a Twitter post, which people ask me, how much did this cost?

This study cost roughly 20 bottles of wine, 50 Starbucks gift cards, and then my effort. And so, what I would say is that it's a nominal amount of money to do these. Priceless. This is a MasterCard ad or a Visa ad. That's amazing. That's amazing. Okay. So, every time there's sort of a—. Just to get the order of operations there right.

Did you get buy in after the 20 bottles of wine had been consumed, or was that a thank [00:17:00] you gift? That was a thank you gift. And I would say that was not paid for through my startup. That was a personal thank you gift. Got it. Got it. That's amazing. You know, so maybe just walk us through a little bit about the study in a little more detail.

So, what was the sort of prospective component? What was the evaluation, the sort of figure of merit, what were you evaluating? And it's kind of similar to the question that I asked about the first paper that we talked about, what was the magnitude of the difference between the two arms here, between the AI read and the sonographer, and how did you tell that that was clinically or potentially clinically significant?

Yeah. Echo is a really unique place where there are two clinicians looking at every individual image, by which I mean that the sonographers or the ultrasound technicians that acquire the images oftentimes do a preliminary assessment as part of standard clinical practice, but they themselves are very much experts.

So, in the standard clinical practice, a sonographer gets the images from the patient, actually traces [00:18:00] the left ventricle, and then have a preliminary assessment of ejection fraction that's finalized by a cardiologist. Because this is a two-clinician task, we thought this was a perfect place to blind and randomize, meaning that we can inject the AI model to duplicate or simulate the work or part of the work that a sonographer does.

And then because it's already done asynchronously, the cardiologists have a set of echoes that they have to look at. And if you shuffle and randomize on any individual study, they might not know whether it was done by AI or sonographer. Our primary endpoint was the difference in the preliminary assessment.

So, whether by AI or sonographer, as well as the final assessment of that same study that's been adjusted and finalized by the cardiologist. We thought a more than 5% change in ejection fraction was a clinically meaningful change and we were looking at it initially as a non-inferiority study but assessing for what is the proportion that was changed by the sonographer arm [00:19:00] and the AI arm.

Can, maybe just to, to ask a follow-up question there. One, I mean, I'm sure because they were enrolled in a study, but were the cardiologists aware that they could be grading the output of an AI system? And if so, did you capture their self-perception of whether or not it was from a sonographer or an AI system?

Yeah, the cardiologists were aware that they were participating in a trial. Any individual study, they did not know whether it was a sonographer or AI. And after each study, we actually asked them if they thought it was AI or a sonographer. So even outside of the metrics of precision and accuracy that we assessed, we evaluated how often were the cardiologists correct.

They were correct about a third of the time. They were incorrect a fourth of the time, and they just didn't give us an answer or said they can't tell the rest of the time. Were there any interesting, like sub-analyses of the ones that they were correct? Did they have any inherent bias? If they thought it was an AI, [00:20:00] did they grade it differently than if they thought it was from a human?

Yeah, there was actually minimal subgroup variation, whether they thought it was AI or stenographer, they changed it less frequently when it was the AI arm. Or there was less change in initial to final EF in the AI arm in almost all subgroups. Because I could just imagine this like complex cognitive bias system where they're not actually grading the thing.

They're trying to guess who made it. And depending on their own cognitive biases, if they guess that it was an AI derived thing, then they mark it up. Or if they guess if it's a sonographer-based, then they don't. But it sounds like that that didn't really come through in this study. Yeah, it's a really good point and a big reason why we think blinding is really helpful and necessary for the assessment but there was no subgroup variation.

Cool. Awesome. And David, you said that the first author of the paper, Brian Hie, that he had to convert the images or convert the videos or convert the outputs, right, [00:21:00] into the sort of the same kind of standard way that the cardiologists were used to reading these studies, right? Yeah, so one of the benefits of AI is that AI doesn't tire.

It can actually measure every single frame and every single beat, but that would remove the blinding. So, we intentionally weakened the AI and Brian actually reverse engineered the PAC system and did a SQL injection so that it actually looks exactly like and was annotated in the exact same way as if a manual annotation was done.

So, 20 bottles of wine, something else, and one SQL injection to accomplish this RCT. I was about to say, this is the only time the hospital IT staff has heard SQL injection and not immediately had a heart attack themselves. Right, right. So, this is very much, I think, hospital IT was very helpful. We did this all on the development branch of our PAC system.

They were definitely and appropriately skittish of us doing this in the production branch, but even having access to the development branch, is something that, Raj, to your initial point, needs a lot of soft [00:22:00] power and, needs a lot of coaxing and encouragement and collaboration. So, I was going to ask a follow-on question about where this work is at now, but I think this might actually be a good transition point to what Andy wants to talk about, which is how you start companies and how you move from academia into, into products.

So, Andy, do you want to take it away? Yeah. You read my mind, Raj. So, I think this is, so when I like sometimes, like we'll teach to or lecture to folks from the business community about how you develop an AI product. And literally mainly what I'm doing, David, is telling them about your work. Because not only have you done the like initial model development, the trial, you then have also passed the academic membrane and gotten FDA clearance and have a commercially available product.

So, could you tell us the legacy of these several studies that you've done, what your commercialization strategy and journey has been like? Maybe we'll pull out some threads there that could apply to lots of folks who are listening. Yeah, Andy, this is a really great question and something I'm really passionate about.

In 2022, [00:23:00] after the RCT was done, Brian and I thought this is really the right time for clinical deployment. This is not necessarily the skill set of an academic medical center. In fact, there was a lot of regulatory hurdles and efforts necessary to get FDA clearance. But we underwent the summer batch of Y Combinator, accelerator.

We fundraised and we thought this was really the perfect place to deploy AI. I would say that our company, InVision Medical Technology Corporation, has

the same thesis as my lab, which is that we think that AI can help the accuracy of cardiovascular diagnosis. It will both improve the lives of patients and clinicians, and that we have a unique perspective on how to get that into clinician and patient hands.

Awesome. So, could you say a little bit more about your YC experience? So, for listeners, YC is Y Combinator. It's this very well-known startup accelerator where you go in almost pre-product, and they help you develop, they help [00:24:00] you accelerate the idea, and eventually lead with some seed funding. So, can you give us a little peek inside like what the Y Combinator machine is like.

Yeah, the great part of co-founding and leading a startup is that there's a lot of learnings that I also take in from my research lab itself. The first is to work hard and work fast. So, I would say that when we did Y Combinator, we were very much encouraged to really have something ready for demo day to really build relationships with hospitals that I really helped start some of those conversations, and really to think about what is the next stage, actually just work harder.

The bias towards action, as people like to call it, is in of itself a superpower or good reason why startups are really great and potentially have a competitive advantage compared to incumbents. But we really felt like the opportunity was there and it was an opportunity that I think will inevitably help patients.

I thought I was listening to Paul Graham for a second there. That was great, David. [00:25:00] So now putting on my entrepreneurial hat, who's buying this? What is the business case? What is the market? And what's the commercialization strategy for, if I'm a hospital CIO, why do I want to, what's the use case for me?

I'm putting money down on this. Yeah, Andy, that's a great question. Fundamentally, the biggest challenge for software as medical device companies right now has been reimbursement or what is the financial value versus what is the clinical value. I think through our papers, we show that there's strong clinical value and proves the precision of this assessment, which has a lot of downstream implications and diagnosis and treatment of patients.

And it's also something that saves sonographers and cardiologists time, which has business value. The challenge in this space is that oftentimes AI algorithms that reproduce or aid clinicians in the assessment of things that are standard assessments do not have a strong reimbursement strategy. I would say CMS as

well as private payers will say actually that's something the cardiologists already should do.

Why should [00:26:00] we give you additional payment to do this? This is an area where we have to pitch the idea that this very much streamlines and improves clinical value. But potentially for a hospital CIO or CTO that we need both the clinical use case and the clinical value as well as the business value where that becomes a little bit weaker.

In parallel, we've really focused on disease diagnosis, particularly we have an FDA breakthrough designation algorithm for the screening of cardiac amyloid, where we found much more alignment between all the stakeholders. Cardiac amyloid is a rare disease. It's often missed, but the necessary information is already in the echo.

So, we've built an algorithm that was a *JAMA Cardiology* 2022 paper that actually allows for early screening and early recognition of this disease. This is something that we've actually collaborated with life science partners, as well as I say nonprofit organizations that are interested in improving the precision of medical diagnosis to really push a product like that forward.

We recently [00:27:00] got a CPT code for this that will actually be available in January of 2025, and that is where there is alignment because this is a task that clinicians traditionally are not very good at. That's awesome. So that's the second CPT code, I think, on the podcast, Raj. Abramoff was on the podcast a while back.

He got a CPT code for his device. So, I think that makes total sense. David, I wonder if I could get your thoughts on the cognitive dissonance sometimes in medical AI, specifically as it relates to commercialization. So, I think without a doubt, you, as you said, you have demonstrated the potential clinical and patient benefit of this technology. It works tirelessly.

It never gets tired. It never goes on vacation. It's as accurate as the human equivalent. And yet there's this challenging value proposition that we have to make. How do we fix that? Like, cause that to me has always been one of, and I know that's like a very broad and big question. We had Vijay Pande on a couple weeks ago and asked him like a similar question. But like, this [00:28:00] to me has always been like some of the tension, in medical applications for AI is that you can actually have the world's best product but there may not be an actual product market fit for that because of these like competing incentive structures.

Yeah, no, this is a really great question, and I think that a lot of companies in this space are struggling with that. I can't say that I have the solution. The hope is that if you create great clinical value, the business value will come. But definitely not having an M.B.A. and not necessarily having a large war chest to actually push forward a lot of kind of regulatory or kind of financial changes.

The real hope is that eventually more and more people will use the product. We'll see that it's valuable. And then there's more of a business case for its value. I wish I had a better answer. I think that's the best answer that anyone can give right now. I was just curious, given your lived experience of being on both the founder side of this and the clinician side of this, but I agree.

It seems intractable. If I knew how to solve it, I [00:29:00] probably wouldn't be behind this microphone right now, because it would be a huge opportunity. This is where I would also highlight our relationship with life science for our cardiac amyloid product, which is, that is actually an area where there is more alignment, by which that in the last five years or so, there are multiple new therapeutics that are targeted for this disease that both improve morbidity and decrease mortality.

In fact, one of the ESC headliners for is another therapeutic, but ultimately this is an area where there's alignment because it's an underdiagnosed disease that has a really potent or multiple potent therapeutics. And so, would this be like a, like a companion diagnostic kind of thing where there's a diagnosis that then you would need for treatment?

It's an area that's tough because you also don't want to self-deal. There are laws against having companion diagnostics that directly funnel to a particular company, but this is an area where there's a clear unmet need and there's a clear therapeutic. One big challenge for AI is that, almost [00:30:00] uniformly, AI is a diagnostic where a lot of the ways that we think about medical treatments are for therapeutics and we try to pigeonhole a lot of the metrics for, into kind of the framework for therapeutics.

Got it. Yeah. And reimbursement always happens for therapies, which makes those business models like much clearer, which is why you see like huge rounds, uh, for biotech companies making therapeutics, but diagnostics have this kind of imbalance. So, thanks, David. I think that was fantastic. I think this is a great time to transition to the lightning round.

If you're ready. Let's do it. You have so far done what few have done on this podcast, which is escape LLM discussion this deep into the episode. So, we

have to bring it up now since it's been such a popular topic. We've heard a variety of opinions on this specific question, but I'm curious on your take given your like very cross-functional background.

What impact do you think LLMs will have on medicine in the next five years? [00:31:00] I think LLMs will have a tremendous impact, and we're already seeing it, both kind of the submissions that we're getting for *NEJM AI*, as well as kind of use cases from companies ranging from Epic to Abridge and Microsoft. That said, maybe this is a controversial opinion, but with trainees, I don't recommend they do LM based research projects in my lab.

The reason I say this is that for any research project, you both want to care about the research project as well as you want to learn a reproducible skill that will inevitably be useful. The challenge with a lot of these I would say closed-source LLMs is that you're primarily doing prompt engineering.

There's no theoretical basis or empirical evidence that your strategies for GPT-2 works for GPT-3, will work for GPT-5, and etc. So, I don't believe that that is a reproducible or kind of continuing skill. So obviously it's really useful to use LMs for projects, but I don't recommend it to be the core of any particular focus.

This is also [00:32:00] different if you're using LLaMA or you're actually doing the hard work of fine tuning and training. That is an interesting take because like my naïve thoughts on this is that it actually is a boon to like busy medical students. They can do it independently. So, if you don't have coding skill, if you don't have access to a wet lab, there's still a set of interesting yet transient questions that you could sort of self-investigate.

But I agree if your core focus is skill development as a researcher. Then if you do it in LLMs, it's a skill that will require constant maintenance. Because the, they change sort of like on a monthly, if not weekly basis, what you did for a paper now, it's probably not helpful or relevant for what you might have to do a year from now.

Yeah, I would definitely say for non-technical members of the lab, say potentially someone with a medical background that doesn't have any prior computational background, it might be a reasonable project. But my main hesitation [00:33:00] is that oftentimes, it's like cotton candy. You put a lot of effort into prompt engineering.

You think that you have a really robust thing, but then it becomes really brittle, or the more you touch it, you realize it's more of an assessment of the model or assessment of your prompt engineering strategy than an underlying theme that's related to medicine. I agree. And in honoring the spirit of the lightning round, I will hand it off to Raj for the next question.

I was going to say, I was like, that was the, it was so interesting that I didn't want to break it up, but it was the longest lightning round response ever. So, congratulations on that. I think, yeah, this is not lightning. I think we discovered that if we don't talk about LLMs before the lightning round, the lightning round turns into a longer, a longer back and forth.

Okay. Anyway, moving to the next question, David, if you weren't in medicine, what job would you be doing? I think I have the perfect job. I would say that it's a cross functional job that does both research and clinical work. If I weren't in medicine, I think I'd like [00:34:00] to be in tech in some fashion. My wife is a product manager.

That seems like a lot of fun. Being an engineering manager of some sort is also really cool. Excellent. Awesome. So, this is a question that much ink has already been spilled about, but I have to ask you, given your background, do you think medical students should be required to take more statistics courses?

I would definitely say, like going back to that *Arrival* theme, that it really impacts and influences how I think. I don't think every clinician needs it, but if you want to be a clinician in AI, it is absolutely necessary. I'm just reliving very long debates that Andy and I would have as post docs sitting at, Countway Library about this topic and we would, it would be like hours and then we'd return back to our, what we're supposed to be doing. We would return back to playing frisbee golf on the fourth floor of Countway.

Yeah, that's what our postdoc was in frisbee golf. Alright, next question. What is an example of a frivolous thing you do or something [00:35:00] you do just for fun? What's your hobby? I like to fly drones. I think it's really fun and it has a different perspective. So, I have actually, my wife complains, but I have three drones.

Wow. See, it kind of relates. I mean, imaging, right? It's computer vision relevant there too, but wow. Super cool. It's like playing video games in real life. Yeah. Amazing. Amazing. Okay. So, this is a popular one that has been revealing. So, we'll ask it again. If you could have dinner with one person alive or dead, who would it be?

Oh, it's interesting. Can it be in the future, or does it matter the age of the person? Again, you would, this would be a first there. The future is within scope, and I'm interested to see what that would be. Yeah. So, I, I have a one-and-a-half-year-old son. I think it'd be great if I could say, have dinner with him where he's my age now and have a conversation.

I think that'll be really fun. And if it was in the [00:36:00] past, you know, similarly my dad when he was younger, or other people that I know well in this context, but hopefully have a different perspective at a different age. And just to clarify that you're not revealing some news here, you mean you'd like to have dinner with your son in the future and you both at the current age that you are now, not just having dinner with your son 30 years from now.

Yes. Yes. Hopefully, I will still be able to have dinner when I am 30 years older. Our last lightning round question. Will AI in medicine be driven more by computer scientists or clinicians? That's like asking, in a bicycle, what's more important, the front wheel or the back wheel? I would definitely say it's both, and I think it's neither is sufficient without the other.

Just keeping up the analogy more, when you're going downhill, you have to know which one to stop first, you know, so you don't, you don't get in trouble, but I love it. Alright, David, congratulations, you've survived the lightning round. We just have a, you know, a few big picture kind of concluding questions [00:37:00] here.

That we want to wrap up with. So, I think we talked about this a little bit with some of the papers that we talked about earlier. But I think you've done something that really is unique. And I really want to emphasize this because I think there's a lot of amazing work that's happening. You know, we're evaluating a lot of great studies at *NEJM AI*, tremendous number of submissions that are very interesting. But what I think we've noticed, and we've been discussing, and we discussed this actually yesterday, right, at our weekly editorial meeting. There is this gap that we notice between, I think, sometimes very interesting technical papers and rigorous clinical evaluation usually in the form of clinical trials, RCTs, that are well conducted pre-registered where there's an outcome that is very clearly relevant to what the authors are trying to assess and hopefully clinically relevant or clinically relevant as well.

And I think they're, it really is a skill, right? And, you know, we were talking about this yesterday. It seems like something that, if you just look at RCTs, it should be easy, right? [00:38:00] Have one intervention group and you have

like one control group and then randomization takes care of all this sort of confounding that you don't really understand about the world.

And you end up with a proportion difference or some other outcome that you're looking at, but there's a lot of nuance. And it's very difficult just to formulate the question correctly. And then as we talked about, it's also sort of socio technically challenging as well. We have to either raise funding or wine bottles and SQL injections or other things to raise funding.

You have to convince people that this is meaningful and important to do. That it's ethical to do. And then you have to convince providers or sites that are then enrolling patients, and then provide informed consent, and enroll patients. There's so many sorts of challenges and important steps along the way.

And so, I, I think it's no surprise then that there's a lot of amazing research that's not RCTs and that there is a real dearth and need especially medical AI, but I think across medicine, right? For an increase in robust evidence, typically in the form of [00:39:00] RCTs to know what works and what doesn't work.

And you've done something amazing, which is, you know, you figured it out. And I think you give us some clues there, which is that you're creative. I think you did this before you had, you didn't have $10 million from a company to support what you were doing at the time, right? You were able to bootstrap it and be creative about how you launched it.

But how did you learn to run a clinical trial? Like what is for other researchers, especially other medical AI, machine learning researchers, what should they do if they want to run a clinical trial? What should they learn? Who should they talk to? What skills should they try to acquire to run a clinical trial and to move things from interesting retrospective or even prospective study to a full on RZT or clinical trial to evaluated technology.

Yeah, this is a really interesting question and something that I've thought a lot about. The two ways I would frame it is, one, incentives drive behavior. So, the reason why we're not seeing as many RCTs is [00:40:00] that to get FDA cleared, you don't need an RCT. So, many commercial companies don't find the value in doing an RCT unless they're pushed or necessarily asked to do so.

So, we aren't seeing as many RCT in commercial products. Hence, I would say our EF algorithm, was RCTed before it was FDA cleared or before it was a commercial product. But the second piece is that similarly going back to the

question of incentives driving, I guess, behaviors is that there's the difference between the cost of running a trial and the business of running clinical trials.

So, clinical research organizations are actually a high margin project, right? You've heard very big names, in academic medicine, but they charge an arm and a leg because they can, because their primary customers are pharma companies. Because, you know, quote unquote, the EF algorithm was our baby. We were doing it at cost and at cost

it's a very small proportion. It is a small percentage of the cost of if you actually had to ask a CRO to do [00:41:00] this. I would say that the teachings that I've gotten through InVision and running a startup is you just got to do it. It is, there is no gatekeeping. There is the more you try, the more you'll learn through the process and the better you'll be because of it.

As a clinician, even going through residency and fellowship, you already see many clinical trials. So, I think the chair of medicine at Stanford was previously Bob Harrington. He was a trialist. And when, every time we meet him, we learn something, a new wrinkle or interesting idea about a clinical trial.

But by purely being a clinician and in the field, you should already have some intuition of what you like or don't like about trials. The second piece is, I would say, to the first approximation, most AI technologies are diagnostics, so it should be much cheaper. The turnaround time, the outcomes, all those things could potentially be much more pragmatic and, much more, uh,

I think, short timeframe that allows for more iteration. So, I think a lot of it is, [00:42:00] I would encourage more clinicians to really just do and to push forward. And I think this is an area that kind of Chai and many other places are trying to figure out a way to actually advocate for more prospective evaluation.

So, we had MasterCard earlier, but this is Nike now. Just do it. Right. Yeah, you know, you imagine this being, I think it's very easy to imagine this being something that is for someone else to do, with decades more experience or who has much, much more in the way of resources to actually get off the ground. But I think you're at least an existence proof, but you're much more, but you're an existence proof that you could do this very early, very young and without significant resources to, to launch a very interesting and impactful study.

I think that's like such an interesting insight you had there, David, too. Cause frankly, as a non-trialist, I would feel completely intimidated and overwhelmed because of the perception of how expensive and complicated they are. But you

made this like very, very good point is it's expensive and [00:43:00] complicated, or at least expensive because of who is running it.

Like who wants the answer. And these are usually organizations that have lots of money that are like well capitalized and can tolerate an expensive high margin trial. Because there could be a blockbuster drug on the other side of it, but you can actually do it. You can actually do, you know, the equivalent of like a lean startup or a, you know, a lean trial, especially in AI, because there are lots of things that are favorable from a cost and time perspective that are not true

and like a big RCT of a new, drug. So, I think that that's like, a very, very useful piece of nugget and a place where clinicians can have extreme ownership and like really build like a very clinically important brand around clinical trials of AI technology. Alright, so I think we're going to move to the last big picture question.

This is I think firmly in the category of Andy-style questions, which are less about clinical impact and more about wild-eyed [00:44:00] futuristic types of concerns. And so, I'm just like, especially given your geographic location on the West Coast, you're a YC alum you've worked with Ewan who also has like lots of startup and entrepreneurial things.

The vibe that we on the East Coast get from the West Coast is that the future is going to be here any second. That GPT-5 is going to take over the world. We should all have like our fallout shelters for when AI takes over the world and that the rapture of the nerds is here.

Repent. Is kind of the vibe that, comes from certain parts of San Francisco and the Valley. I guess given that you're West Coast and in cardiology, what is your trajectory, or what do you think cardiology will look like in 10 years? Is it, like, a more efficient practice enabled by AI?

Is it different ways of practicing entirely, or is it something that's just very, very hard to predict? Yeah, Andy, that is a very tough [00:45:00] and important question. I think that there are maybe two pieces that I would highlight on a high level and why I think I'm really happy with my job as it is. The first is that medical AI will inevitably need people like the three of us, because by definition of the domain and the type of data, most large scale LMs do not have access to this.

By virtue of being a clinician, I have more perspective on what is good data, where is the variability, and what are the pitfalls in the data, that I think that it

would be really hard to have a pure nonclinical or LLM, to approach. So, I think that that last mile deployment is so important that inevitably, I think to the first approximation,

it'll still look relatively similar even in 10 years for the clinical practice of cardiology. One analogy I like to give is that [00:46:00] the ChatBot UI is an artifact of LLMs. It's a bug, not a feature. If you're trying to learn tennis, you probably would not learn it well if you're just using LLMs. And in the same way, if you were trying to practice medicine through ChatBot, that's not what clinicians want and that's not what cardiologists want.

I think that it ends up being things like our algorithm that is actually directly deployed in the clinical packs or the clinical EHR, all these things that are directly integrated that has more of a meaningful impact. I was listening to a podcast recently with Nat Friedman, who's the CEO of GitHub.

And he actually said that for GitHub Copilot, the adoption and utilization really skyrocketed when for Copilot, the interface was essentially optimized. Their initial iterations were, why don't we just give you a huge code chunk, and then there's multiple choices and you can choose, but that's like really a clunky way to actually deliver information, even if you have a great model. [00:47:00]

So, in the same way, clinicians, I think, I personally wouldn't want to have to keep reaching for a chatbot to answer a simple question. But if it was embedded into my EHR, embedded into my PACS, it is actually allowing for more streamlined and efficient care. So, I guess I'm asking less about do people want to use bad products and more about your conditional probability that the thing that you actually want to use happens and changes practice over the next 10 years?

Yeah, I, I might be a pessimist, but I think that clinical practice will still look 90% the same, 10% more improved. Right. Uh, and I would go to say that it is, we often forget that maybe a lot of the inputs doesn't, isn't, or it doesn't have all the necessary information. Even if you have the best decision making, if there's uncertainty, or if there's the lack of information in that input modality, you can only improve it so much.

So, I would say that [00:48:00] kind of, I still see cardiologists seeing patients. I see that potentially we will assist them with EKGs and echoes that can flag things that they're not sure about, but at the end of the day still reflux into the gold standard therapeutics that are necessary. Er, diagnostics and therapeutics. Okay, so that's helpful.

Let me then ask you another sort of statistical estimation question. What fraction of diagnostics in cardiology does AI do mostly autonomously over a 10-year time horizon? 80% more likely than not. Yes. That's a pretty sizable chunk. That's a pretty, we didn't have a chance to chat about it, but we, we like how EchoCLIP looks, we have new versions of kind of video language models that very much simulate the task of interpreting images.

So then, okay, this will be the final question. What is the David Ouyang as a cardiology fellow [00:49:00] in the year 2034 learning to do in cardiac fellowship if 80% of the diagnosis has been taken over by AI? SQL injection.

I have never regretted learning a field deeper, whether that's medicine or kind of computational. Like, if it's hard for you, it's hard for everyone, and it's a moat or a defendable skill. I hope that the future cardiology fellows still take time to really think about what they're doing and when algorithms or tools are incorrect or how to evaluate quote unquote bad data, and I think that will be persistent across time.

Awesome. Critical thinking never goes out of style. I think everyone has to agree with that. I love it. David, thank you so much. That was amazing. Andy, Raj, this was really fun. Thanks for inviting me. Yeah, thanks for coming on David.