AI Grand Rounds Podcast #3 02.14.23

NEJM AI: Dr. Lily Peng: AI for Ophthalmology and the Challenges of AI in the Real World

[00:00:00] Welcome to the third episode of N E J M AI Grand Rounds. Today we are thrilled to bring you our conversation with Dr. Lily Peng. Lily is a physician scientist and director of product management at Verily and Andy, this was a really fun and enlightening conversation for me. I learned a lot from Lily about both the real world challenges of developing and deploying robust medical AI from her team's Landmark paper on detecting diabetic retinopathy from retinal fundus photographs to some of the stories that she shared about their efforts to deploy those AI technologies in Thailand.

She really illustrated the importance of this thing that we keep talking about over and over again, uh, which is the data generative process. So what is embedded in our data and the labels that are used to train our machine learning models can be subtle, but it can really [00:01:00] determine success. Lily also highlighted, I think one of the best antidotes for addressing this, uh, which is talking to and partnering with clinicians who can teach you about the assumptions that are baked into your data.

I couldn't agree more. I've been a big fan of Lily's for a long time and the paper that you alluded to, the one that came out in 2016 was definitely a "I remember where I was" kind of moment, because for me it was one of the first times where a good quality data set, along with a really important question, came together with emerging AI technology and was just such a compelling demonstration for all of us of what was possible.

But what impresses me most about Lily is that she and her team did not stop there. They actually took this technology into the real world and did an honest assessment of where it does and does not work in actual clinical workflows. And that's just such a rare thing to have a single team behind my technological breakthrough, but then also sort of the real world impact.

So as I said, I've been a, a big fan of Lily's for a long time and it was such a pleasure to sit down and [00:02:00] talk with her about what she's worked on in the past, but what she's gonna be working on going forward. And so with that, we're happy to bring you Dr. Lily Peng.

Well, I'd like to welcome you to AI Grand Rounds. Lily, thanks for joining us. Oh, thanks for having me. So, Lily, this is a question that we always get started with. Uh, could you 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?

Oh, wow. So I started getting interested in ai when I was working on search at Google. I had heard about this 20% project, which at Google it's a program that allows you to spend 20% of your time on kind of anything you want. Um, and at that point there was a research team who was trying to train image recognition, uh, algorithms.

Essentially, they were looking for people [00:03:00] who would be able to evaluate images and create labels for those images. . And of course I can't label anything. I don't have the clinical expertise to do that. But as a physician by training, I had friends who were also physicians who could do such a thing. And so I would say it started off being a little bit serendipitous.

I was just volunteering for something that I thought was super cool. And then over a time that became sort of my full-time role at Google because the team just thought that, uh, this was, well, one, that there was a lot of impact here in terms of being able to help folks. And then two, I think it was a project that really captured people's imaginations.

And so that allowed us a bit more breathing room and a lot more, uh, teammates to join us more full-time. . Like all great epics, you started I think in medias res there, like in the middle of things. Uhhuh . Could you take us back a little bit further to how you got in? You're an MD PhD by training. Uh, could you tell us sort of like what sparked that initial interest that landed you at Google in the [00:04:00] first place?

Oh yes. Okay. So like great epics, I'm gonna rewind us maybe 10 years earlier, or something like that. So I started being interested in research in general, actually in high school. So that's more than 10 years earlier and was interested in trying to figure out. How we could take innovations and new things that weren't discovered before and make them useful.

And so that's how I started off with thinking about research. I started in a C Elegans lab, so for nematodes and did some research there, loved the work, but at some point really wanted to think about how this actually impacts patients. And so in college and beyond, I started working more like, you know, inch my way up from nematodes to mice, and then eventually more to humans with medicine.

So I trained as an MD PhD because of that, because I wanted to do that translational work and loving school and school is all I did. Uh, thought that [00:05:00] I should get as much of an education for as long as I could. So I did the MD PhD, did the work in bioengineering. And so that was what set me up for the work that I eventually ended up doing at Google.

Cuz at some point I had to leave academia and also think about how businesses and startups actually contributed to this translation. Cuz it wasn't just academia, it was that academia had to work with industry as well to do the translation. So that's what happened. I started in academia, worked my way through, and worked my way through the process of thinking about translation.

And now ended up more on the industry side. And I'd like to dispute the claim that you have to leave academia. You can stay there for as long as you like . That's true. Um, that's true. I think something interesting about you, and I think I have this correct, is that you, so you didn't do residency after you did the MD PhD, correct?

That is correct. I did not do a residency. I always thought that I would, and so I like to tell myself that [00:06:00] I didn't do a residency because my residency was in product management and I learned a lot from product management and how to help patients through building real products. But if I'm really honest with myself, I feel like I didn't do residency cuz I really like sleeping and you can't sleep during residency.

and . I just couldn't give it up. . Yeah. That type of rebranding, you would've had a promising career as an academic, um, that you did, uh, a residency and you, you really missed your calling there. I think so. So I think Raj, do we wanna jump to the research portion and, and dig in a little? I know that I've told you this to you before, but I'm gonna say it again.

I think that the paper that you wrote on Diabetic Retinopathy is one of the seminal papers in the history of medical AI. I think it was like one of the first examples where AI was taken to a meaningful clinical problem and where the paper was executed flawlessly and really I think set the standard for the next decade of medical AI.

So what I'd like [00:07:00] for you to do is to walk us through this seminal paper. Maybe first tell us a little bit about what diabetic retinopathy is and the clinical problem that you were trying to address, and then how the AI solution addressed that. Oh, awesome. Well, again, thank you so much for saying that. I think when we were.

in the progress of writing the paper, it certainly didn't feel like we were going to publish anything seminal. In fact, we weren't even sure we were gonna publish anything. Interesting. . So, so let me walk through this. The first question around what is diabetic retinopathy other than a mouthful? Diabetic retinopathy is a complication of diabetes that causes blindness, if not caught early enough and treated.

And so the way that you detect diabetic retinopathy is by taking pictures of the back of the eye with a special camera. And in general, clinically we recommend that people get that screening once a year. And so you can imagine with the number of diabetic patients that exist now, this is a very [00:08:00] common procedure and also something that can overwhelm the medical system very quickly cuz everyone needs the screening in particular.

We were working with folks in India that was actually doing the screening. They had a long history of doing screening, uh, to prevent blindness due to cataracts and other things. And they were seeing a growth in the number of people at risk for diabetic eye disease and actually started a program screening diabetic patients for the disease.

So that was how we started. We actually, there was another Googler who was visiting India at the time and met a few ophthalmologists that was doing this work and, um, actually brought home this problem and said, Hey, you know, is there something that we can do on the AI side to help with screening, especially in India?

So that was a origin of the project and it was a really scrappy project because it wasn't something that was a clear business priority or something like that at Google. Right. We, we were in the research team [00:09:00] thinking about this kind of very interesting problem. . And so at that point it was just a, I think a couple of engineers thinking about the problem.

And that's when I got interested in it because they were asking for people to help them generate labels for a data set that was coming from our partners in India to help. So we've come full circle to your origin story. Now we're back. We're back to the future. Um, and this was your 20% project that you mentioned, uh, like 10 minutes ago, I think.

Exactly, exactly. So this was the 20% project. So that was the beginning. At first, we didn't know what we were doing. In fact, the funny thing was the first algorithm that the team was trying to train was an algorithm that detected the presence or absence of diabetic retinopathy. . And then as we talked to the clinicians a little bit more, they were like, that's not useful.

We actually need something that actually discriminates between more than mild diabetic retinopathy, because when you have mild retinopathy, that doesn't really change the way we [00:10:00] manage you just yet. Right. So that was. first lesson learned, which is always, always talk to clinicians and the users who will be using your product or your algorithm or your research, right?

Otherwise you'll end up making predictions that are like academically sort of interesting, but not actually useful. So, so that was the first, um, there's the anti academic bias there. Again, I love it. . Sorry. But, but I actually think that's one of the beautiful things about being in academia though, right?

Your colleagues are actually right there. If you wanted to reach out and talk to them, they're an email away or a few buildings down. So actually that was really great. Our collaborators were across the globe, so it was a little bit harder. I used to take like really late night meetings so that we could talk, right?

So that was the first lesson. I think the second lesson that was interesting was that we thought we would do all this fancy ai and that was what was gonna [00:11:00] make the project work. And it turns out, at the end of the day, it was about data quality, right? Um, it was about getting the labels that actually accurately reflected what was in the picture, and then getting enough of that kind of high quality data actually made the biggest difference.

So those were the two pieces that were kind of lessons learned from that paper. And you'll actually see it in the paper, right? So the paper actually describes, starts off saying, we train the model that it performs really well on par with experts, essentially. But at the end of the day, if you look at the methods and you look at all, all the figures, it's really describing how we generated the ground truth, right?

Mm-hmm. , what does it mean to have moderate diabetic retinopathy? Is it when one doctor says you have it? Is it when three doctors say you have it? What if they disagree? Right. So that was actually for me, the interesting part of that paper and the [00:12:00] learnings that carried us through beyond, Hey, you know, we, we saw that this model does really well with images.

Cuz to be honest, we already kind of knew that deep learning would work really well on images from mm-hmm. from the consumer space. But what we didn't really understand at that point was how to adapt it properly to the medical. . Got it. And just to like put a finer point on some of that, I always have a, a mental image of how these deep ordering systems work and they're sort of like data goes in on the left side, a big sort of neural nets in the middle, and then the annotation that you want it to predict is on the right side.

And if I understand you correctly, what you're saying is that the annotation from the human physicians was both very important but also very noisy. And so your team spent a lot of time getting these consensus annotations for the grade of diabetic retinopathy that each patient had, and that if you didn't sort of stabilize the annotation for each image, the algorithm might not have been able to learn anything at all.

Is that a [00:13:00] correct sort of summary? Yes, that is absolutely correct. And. , I probably don't have to tell you all, but getting doctors to agree on anything, even with themselves, sometimes it's really hard. So we actually started measuring, you know, like integrator agreement. Integrator agreement, so that we actually had a good measure of when disagreement actually happened.

And the interesting thing about this is at some point later on you realize that disagreement in itself is a signal, right? Mm-hmm. and in terms of the complexity of the case. So I think it was important to get, not to just average all the opinions, right? Which I think sometimes, like you, you wanna do, you're like, well, you know, there's eight opinions and four of them agree, said it was, there's no disease.

And four of them said there was really severe disease, so let's just take the difference and say there's moderate disease. Like, that's not how it works, right? This is a hard case. And so it needed, uh, additional love and care and understanding of what's going on. So, . [00:14:00] Got it. And I'm just curious sort of like culturally how this project proposal landed at Google, because I can imagine like you going to leadership or saying and say, Hey, we want to create an AI system that can detect this diabetic condition that happens in people's eyes.

Sort of how did that go over, I mean, 2016 before the sort of wave of medical AI stuff that you kicked off had really gotten going? Yeah. Um, , so I probably shouldn't say this, but at first we just didn't go to leadership really. Right. We just, we thought this was great. And one of the great things about Google is that the company truly supports projects that are believed to be high impact for society overall.

And there's a lot of projects like that. So it's not atypical to start projects that actually, it's pretty clear that there's gonna be high impact when you have to go to leadership is when you need more people and more money to do the work in a more reliable way. . And in that case, we talked to our head of research, made a [00:15:00] case for, you know, why this is interesting.

And, you know, really got no pushback. Everyone understood how important this might be to be able to demonstrate that AI could really benefit a really hard problem in healthcare. And so we didn't really have to pitch it that I think the, the story, quote unquote kind of told itself and everyone wanted to support something like this.

So I started as a 20% on the project. I actually had a full-time gig in search, you know, working that, um, full-time. And my manager was wonderful. That's not atypical at Google. He just gave me free reign to spend on this project. And over time I moved, um, full-time on the project. Awesome. So I think that it's, I I was just gonna put why one more, uh, coda on this is that this 20% project, if memory serves JAMA deemed to be one of the most impactful papers that they published in the decade of 2010 to 2020.

So I think that that's a pretty good return on a, like a 20% kind of project investment, um, that you were able to have this sort of, uh, outsized impact. [00:16:00] Yeah. Yeah. To start off a hundred percent. To start off 20% and be able to get through that, I think was, uh, uh, was a huge, we, we ended up getting way more investment from the rest of leadership after we showed that this was possible and we wanted to make sure that the science was good and reproducible and robust.

So it was a bit more than 20% to get the paper out a lot more. But yeah, that was it. That was not what we had contemplated. We didn't even think, we weren't even sure it was gonna be published. Right. And we, we didn't, um, the funny thing about this paper was we didn't understand where the right home was for this paper because it was computer science.

Uh, it was medicine and it was data science, and we were like, well, where is the right venue for this thing? Right. Is it a, you know, in computer science we have conferences and we go through conferences, but the medical community wouldn't see this and wouldn't be, you know, wouldn't see it at a, at a computer science conference.

And then the [00:17:00] medical community doesn't understand AI and deep learning. Right. And you know, if we were to submit it to a medical journal, we would have so much explaining to do about the basics of how this works and is a medical audience going to understand or appreciate essentially the mathematics that go into this.

So we submitted to JAMA fully expecting them to say, our audience is not interested in this. And we were really surprised when they got back to us and said, Hey, we are actually interested in this. Yeah. This, this was basically a paper at that time with no clear home. Awesome. Which was titled Realtime Diabetic Retinopathy Screening by Deep Learning in a multi-site national screening program, a prospective interventional cohort study.

Uh, so maybe we could just start off with that paper. Could you tell us a little bit about it? So the story continues after we published our paper with our, uh, colleagues in India and also a screening, a large screening program in the us. We were introduced to the person who [00:18:00] ran the screening program in Thailand because, People who run screening programs know each other,

And the folks in India actually knew the person in Thailand. And we actually started to talk about what implementation might look like in a national screening program. And so what we did, rather than, you know, I think some people would think, oh, well just, you know, throw it into the population and then it'll definitely work.

Right? That's not how implementation actually works. So we actually, when we worked with the Rajavithi Hospital, which was actually running the national screening program, or one of the main sites that was part of the national screening program, we actually broke down the implementation in multiple phases.

The first one was to do retrospective validation. So we took existing data, existing screens, and we compared how our model compared to their graders.

So it was actually important to compare to the graders. [00:19:00] In country cuz they may grade it slightly differently. And we wanted to make sure that we were grading it to the way that they actually, uh, thought about diabetic retinopathy.

And this is not uncommon, right? Different countries have different grading scales, or different programs have grading skills and slightly different referral pathways. And so Thailand was no different. So we compared images from all, I think 13 of their health regions with graders from the 13 health regions previously prior to this paper.

And then once that worked, we set up a prospective trial to prospectively enroll people across different sites, across different regions. Not just to look at how the model did in these real world settings, but also understand how that affected the system overall. And what were the attitudes and reactions from both the staff as well as the patients going through it.

One of the things is it's not just [00:20:00] about the model and how or accurately it performs, it's also what hassle and what friction do people experience both getting the images and then feeding it to the model and getting or prediction out, but also what happens afterwards. Right. Like, can you give us an example of some of those hassles and those challenges with integration?

Yeah. Yeah. So, you know, um, there's some places where people who are like, let's, let's just. Talk about, let's start from the beginning. How, how do people get to screening places? Right. So if the screening centers are located too far from where patients live, it's could be really difficult to get people there, right?

If you have to take two days off of work to travel to get a screen, uh, people aren't gonna do it because in many cases this people are only paid for the days and the time they work. And so, you know, it's not actually people being, you know, patients being, you know, irresponsible, don't care about their health.

It's that, look, I have a family [00:21:00] to feed and it's either feed my family or get the screening and I may not even be sick. So, but I know my family actually needs money and food and so there's actually a really trade off there. So presenting to screening in itself is difficult, right? Some screening programs actually provide transportation for things like that.

So that's one thing. Just getting to the clinic, once you get to the clinic is where we actually did a more, you know, uh, more work there in tracking how you go through the clinic is, . Some people go in and, and they're not just there for the screening. They're there for a comprehensive, you know, diabetes care.

So where do you slot in the screening appointment? Right. And how long does it take for them to do the screening this way versus the traditional way? Traditionally what they did before was they would take images of everyone, throw them into a CD, and then ship it off and get them graded all at once by a grader, somewhere in a central hospital.

That was a very typical procedure, but [00:22:00] in this case, if you were gonna get, uh, reading back, you would actually need to spend some time explaining it to the patient, which is great because it's a great way to educate them at the point of care. But it took time and it took people to talk through what the diagnosis actually meant.

Right. Um, so that would add. to the visit and require resources and people trained to do that. And so the nurses would need to want to do that, right? If the counseling was not great or people weren't motivated to do that counseling, then it kind of doesn't matter that you got the result immediately if the result didn't make sense to the patient.

Or let's say you were in a case where you had a severe disease and needed a same day appointment, right? We needed to make sure that there were appointments available. And another tricky case that happened a lot was what if your image was not gradable? Right? Not gradable in the sense, and there are degrees of not gradable as we [00:23:00] found there.

It's not just one or it's just not black or white. It's not a binary gradable, not gradable, right? It's that parts of the image might be not gradable. And if they're not important parts of the image, that's okay if there're important parts of the image, that's not okay. And then also you have to look at a person's a1c and other clinical factors to figure out.

Is this just ungradable, but likely normal patient that doesn't need to be seen right away? Or is it a ungradable picture? But the patient is probably gonna be in a bad spot if we don't get them care right away. So there are all these things that you uncover when you have the privilege of being in that room during data capture, during that appointment where you can do more.

But those degrees of freedom also come with additional costs to the staff on site as well as on the healthcare system. And of course, the last piece is once you find all these people and you're able to identify more people who need care, how do you pay for that? [00:24:00] Right? How most people. We don't intervene soon enough.

They lose their sight. And then you pay for, you know, the disability that occurs, but in this case, you're preventing that from happening. So you're preventing something that's going to happen like way far out. So how do you have enough data to show what is the cost effective intervention, longer term?

So that, that's another piece that I think it was another important consideration as you're thinking about an implementation of ai, but really any other technology. I think you said that there was, the genesis of the actual work on the original work on diabetic retinopathy was that someone at Google happened to be connected to folks in India and they came to the US and you got started working on that problem and then you were introduced to someone in Thailand.

But I'm curious whether this informs your selection of the ideal place to conduct these validation studies. Now, how do you think about [00:25:00] selecting Thailand and India amongst all the other possible places that you could potentially conduct a validation or an interventional study like that you conducted recently in Thailand?

And I'm also curious how the experience has been similar or different in the two different countries in terms of your studies. Yeah. That's a great question. I think so. I mean we've also done conducted studies in the US as well, and we're working, I think through a partner with, with the uk. I think for me, what's been.

more interesting is how the way we pay for healthcare changes the way that technologies get implemented. So when you think about a national care system, that's actually, so the UK and the, and Thailand actually have more similarities because it's actually the government pays for a large portion of the screening.

And they're similar in a couple ways in that like, you know, screening is done actually more in primary care settings, [00:26:00] right? Um, because it's more convenient for patients, they want everyone to be screened and then they've, there's a lot of similar kind of cost effective analysis about how it affects costs overall for the country, for healthcare costs.

So I think that there are some really interesting similarities there. And then India is different. It's actually cash pay for the system. So you actually, the patient generally pays for screening. , uh, we worked with charity hospitals that were really more need space, and so they would also take in folks that couldn't pay and would be brought in that way.

But the coordination was more at a hospital level and where the hospital had connections and that's how, um, screening was done. Uh, and it was hospital, a little bit hospital specific. I think there are national programs that are trying to be spun up right now, but fundamentally the implementation and how that works is sort of different.

Um, and then we started also working with blood labs and other places to, uh, increase access for screening, which [00:27:00] is very new, I think in, in the way that we were thinking about diabetic retinopathy screening in India. . So those are kind of the big differences. And then of course in the US we're insured, not really self-pay, but we also don't have a national system is again, a little bit different in terms of implementation, where you'd have to actually go through screening programs that have been already set up, that service, various organizations.

We're looking at how people access screening. So, and even within the systems, I think there's some big differences in terms of follow up, like how people present to care and how people follow up with that care. So, um, I don't know if that was specific enough. No, that was, that was great. That was very helpful.

With countries like India and Thailand, how do you think about the trade offs of increasing access on one hand to expert medical advice in populations, large populations where this expert advice might be scarce and much a need versus the potential for harm if these [00:28:00] interventions are not rigorously validated and deployed in those populations?

Yeah. This is a really good point actually. So the model that we deployed in India and in Thailand's the same model, we are redeployed in Europe, so you're not retraining it, you're not refitting it, tailoring it to the population. So we do validation to make sure that the model works in that population and that the performance specs are similar to what we're seeing, what we've saw, what we've seen in the European population and before we deploy.

That's true, but we haven't really changed the model itself. It's the same, it's a CE marked model so we don't change, the model itself. Could you just tell our listeners what a CE mark is and, uh oh, that's right. And presumably this is by design that it's, it's not changing. So maybe you could walk us through that too.

Yeah. So a CE mark is, um, basically a clearance to be put on the market in Europe. So it's sort of equivalent to F D A approval. , My [00:29:00] regulatory friends will say that it's not equivalent exactly, but that's what you need that clearance to put it on the market and have it be used. And the regulations change from country to country about what is acceptable in terms of medical device to be marketed.

And CE is, uh, the specific standard for Europe and I guess UK before Brexit, , and now they have their own thing, , uh, that is very similar. And so in India they actually have a slightly different directive, but they do recognize, I think the CE mark in India, um, in Thailand is some, has some similar regulations there.

So this is a really interesting piece in terms of regulatory science where AI meets regulatory science and it's actually a evolving space, right. You know, deep learning is not the first time we, um, try to use machine learning to analyze images. And so if anyone's familiar with the mammography caddy and, and whatnot, so the FDA, um, and other regulators have regulated sort of these image recognition [00:30:00] algorithms before.

And so the regulations there for any kind of software is that you would build something, you would lock it down, you would test it, uh, and then deploy it into the market. And then if you wanted to do updates and, or if you wanted to retrain the model, there's specific procedures that you need so that you can update it.

So it can't be changing all the time because there are all these unintended consequences. Right? So it's called change controls. And so we're no different, right? An AI model is the same as other pieces of software. We actually train it to a certain point. And then that becomes the model that is deployed and it's the same model.

We may do validation to make sure that it works in the population, and we also do updates right to, if we understand that how to make the model better. But those are very thought out and we file requests to change it with notified bodies before we actually change it. And when you change it, [00:31:00] are you allowed to reuse the often laboriously collective validation data that was used in validating the regional model?

Are you allowed to reuse that data? This is a problem that I've become very interested in personally, which is how much we can reuse data for the purpose of re-validating slight to more involved changes to the central model. Of course, there is a risk of overfitting at some point, and this depends on the parameters of the problem, but how do you approach that in terms of using or selecting what data can be used in re-validating a a slightly or maybe more substantively changed model?

Yeah, I think it, it here. I'm gonna give you the magical, it depends answered, right? Yeah. So, right. I think right. , I think it's a discussion with the agency about that. Um, or, you know, either FDA or other notified bodies about what you can use. Um, obviously, you know, getting a newly sourced data is, is the [00:32:00] surefire way to make sure that you can update something.

But to your point, can we reuse that validation data? I think it, I think this is a quickly, people are thinking about this in, in terms of how many times, is it just once? Is it twice? When are you gonna start overfitting? And I think the best thing for us, this is just my opinion, only to help regulators and other folks get to a decision, is really.

provide data for when overfitting actually does happen and when is it safe? I, I think, you know, the agency regulators, they really respond very well to data. That's what they look for. And, and they actually look to the research group to actually look at or to demonstrate evidence of when could be harmful to patients because we're overfitting.

So I don't know that we have great data there yet, or at least granular enough to make a call, but I think that's a very important question. So, Lily, before we leave the Thailand study, could you [00:33:00] give us a summary of the key takeaways that you learned from your experience there? I think the main finding from that study was that the AI performed well in these populations likely due to all the other validation that was done before it.

And I think we found some interesting bottlenecks in the workflow to then inform a larger study and inform the national deployment. Another piece I think that's interesting from studies like that is to understand the potential inputs into cost effectiveness models and whatnot. If you're trying to model cost effectiveness about what is actual performance and what is the percentage of unables and that sort of thing.

So, you know, all these things actually go and inform the additional validation that is required and also the additional analyses that could be useful for deciding how to implement something. So Lily, I wanna switch gears a bit and ask you about some of your other machine [00:34:00] learning papers. So one that caught my eye, I think is emblematic.

Of this growing application of deep learning to identify maybe signals that we didn't think were able to be identified in common sources of data. And so you published a paper a few years ago, which is called Prediction of Cardiovascular Risk Factors from Retinal Fundus Photographs via Deep Learning.

And so as I understand it, you used retinal fundus photographs to identify a number of common risk factors that are used in cardiovascular risk prediction. So everything from age to gender to smoking status. And I'm curious, first if you could just maybe give us a quick summary of that paper. Second, just to ask you a question that we also posed to Euan Ashley a few weeks ago, does this lend itself to maybe this future where we have pan diagnostic non-invasive modalities that tell us about a number of potential disease risk indications that we otherwise don't have?[00:35:00]

I love this right behind, this cardiovascular risk factor paper as well. It was kind of a another 20% slash intern project where we had, um, so it seems that we need to do more 20% projects, uh, we all need to do more 20% projects cuz a number of huge things have been launched from 20% projects at Google.

Yeah, no, I think, but, but that's, you know, that's where creativity, like, I think flourishes is that Yeah. You don't have to do it. Right. And so you're, you're thinking about this. And so what had happened with this particular project was we had someone who was an intern and try to learn how to use TensorFlow and, and whatnot.

And so we said, well, you know, what are the labels that are available in the data? Uh, that we could use as positive and negative controls to make sure that they were, they were training the models, right? And so we said, well, you know, positive control, like probably Dr, you know, , that kind of stuff that we knew about eye diseases.

And then, [00:36:00] oh, well, maybe, we can use like self-reported sacs as something that you can't predict. Uh, . It's a negative control. It's a negative control. And then, and we also did things like, can you predict happiness on a questionnaire and whatnot. And so she trained a model and then she kind of show the results and she was like, oh, yeah, you know, and I can predict self-reported sex.

And I was like, no, you can't. No, you can't. You know? Um, And that, that must be wrong. And so we went through and we're like, oh my gosh. Like you really can. And at a AUC that was extremely high and extremely accurate. Um, and then of course we couldn't predict happiness, self-reported happiness. We could not predict that.

So that was actually the negative control. I thought the eyes were the window to the soul. I thought that you're supposed to be able to intuit those kinds of things. Yeah. Well, if that's true, then the soul does not have record of happiness. . Uh, so . So anyway, that was, uh, again, lesson learned that you'll, you'll find all these interesting things from deep learning that you, you [00:37:00] wouldn't have thought of before.

And then since then we've done more work in trying to figure out, well, how is it seen, you know, predicting self-reported sex. And turns out there's some parts of the literature that talks about the thickness of the macular being potentially different in men and women and whatnot. Um, but what's interesting is that in this paper we showed that.

For some reason it seemed like the AI could actually figure out how to understand small effect sizes and then put that together to then make a fairly accurate prediction using features that nudge the probability one way or another. Human beings, I think it is, we're really great at detecting features with large effect sizes, right?

And making some conclusions there. But then I think the way to combine all the different features and how the relative contributions I think is a little bit harder for us. And so it was interesting how efficiently that the machine learning model could do that. So I think that for me was sort of the big [00:38:00] takeaway for this paper.

Rather than, oh we have like a self-reported sex predictor or whatever cuz like there are easier ways to find a person's gender. You can like ask the patient and all these great things. And also with smoking, right? You can also ask about smoking history. So in itself it was an interesting, a very interesting, I think scientific.

Concept, but the paper itself was, I think the predictions that we were making, I think weren't as clinically relevant as I hope it will become. in the future. So I think we opened a can of worms, , which yep. Maybe good worms that we want around. And I'd like to follow up on that a little bit because I was astounded by this paper that you could show pictures of people's retinas to deep warning and it could predict someone's a1c, their bmi, their smoking status.

But then I kind of vacillate between being in awe and being terrified that the rest of the literature is actually keying in on these secret factors that are [00:39:00] associated with the outcome, but aren't actually the true disease pathology. So how do you think about essentially the deep learning models, we would call them shortcut features.

Mm-hmm. in deep learning, essentially just learning a bunch of shortcut features that are correlated with B M I age and gender that are known risk factors for other things versus what we hope what it's doing, which is learning true disease pathology. How do you sort of like square that circle? Ah, well, , I think that square will never be a circle or that circle will never be a square.

Yes. That circle, I think they'll always have these little edges because, so one thing that we definitely do is do baseline models, right? So we say, Hey, based on just BMI, just smoking status, just these things, how accurate can we get in terms of predicting whatever outcome of interest, right? So we always do these baselines in the same data set to make sure that whatever we're adding with the image is actually additive, right?

That, that there's actual, additional [00:40:00] power there. So that, that's one thing that we try to do to at least tease apart the two, um, conversations. You have to know that those things are potential confounders in the first place, though, to include them in a baseline model, right? Yes, yes. That's exactly the problem is that you do have to know, right?

So we do some basics, but we, we are like, I make up, right? Like for some, like this is the back of the eye, but let's say we do front of the eye things, right? So, so we have a new paper where we looked at the front of the eye and said, Hey, you know, we can predict things from potentially just the, the external part of the eye and eye makeup or not, right?

Can, can be a confounder and, and you don't adjust for eye makeup automatically. You do demographics. So to your point, I think we do some basic adjustments, but we can't adjust for everything, right? Another thing that we try to do is we try to work on explanation of the prediction itself, right? So [00:41:00] whether it is heat mapping or so, try trying to figure out the relative contributions of different parts of the image or a new technique where we actually try to figure out like, which, uh, we, we will.

Dial. The feature, how do I say this? So like, I think we use, I think it's GANS, so we use GANS to figure out, uh, if we were to dial, um, a dial one way to make it look more female or more predictive of one thing. Mm-hmm. , these are counterfactual explanations, I think, right? Counterfactual explanations.

Yes. Yeah. Yes. So, so we do, we go through a series of explanation techniques to better figure out what we think is contributing to the prediction and then changing that to blocking out of image or, and what have you, to make sure that that feature is truly contributing what we think is to the prediction.

So those are a few ways, but. Again, if it's a lot of small effect sizes coming together, that can be pretty hard [00:42:00] to tease out, right? So I don't think we have the right answer to all of it, but there are some basic ways that we can try to square the circle, if you will. Okay. Given where you landed. I think that that's a perfect segue to our next section.

So the next section is the lightning round. Ooh. Um, the rules are, is we're gonna ask you a very short question that can be answered with yes no, or essentially a single choice, and we want you to justify your answer in a couple sentences. We used to impose character limits on the responses, but we damn have since relaxed that

Um, but keep your answers brief and concise, um, if possible. All right? Okay. All right. First, uh, should AI be explainable? Yes. Do I need to explanation? No. Do you need to explain that? . I'm not AI. Uh, that's a little, that's a little meta, but I think, yeah, a little, a little [00:43:00] justification for your, non explanation.

A justification. I think AI needs to work with humans and humans relate to each other through explanation. And so if I expect my colleague to treat me with respect and explain things to me, I expect the AI tool that I used to do the same. We're gonna come back to that, but in the, in the interest of the lightning round, we're gonna keep moving.

Okay. And so these, these are kind of all over the map in terms of, of topics. So Lily Peng, what is your favorite novel? My favorite novel fiction you mean? I think that, yeah. No, we're gonna stay with fiction. Yeah. Harry Potter. Nice. Because I believe in magic. Of the seven. Do you have a favorite?

I like the innocence of the first one. . It gets dark after a while, but, um, yeah, teenage years get a little emo, so, yeah. Yeah. Lily will machine learning reduce costs in healthcare? Hmm. [00:44:00] That has a good question. Oh, man. the optimist in me says yes. Long term. Yes. So we have seen models that are very good with language and text, the most prominent of which are ChatGPT, that can generate from scratch very realistic looking descriptions and things like that.

So the question is, will doctors still be writing notes by hand in five years, or will things like ChatGPT have freed them from that? Well, we, doctors don't write notes by hand. Do they? ? Well, they're like templated, but they're not very good. Will there be abstractive summarization by something like ChatGPT that removes a lot of the grunt work of forming up the patient's record?

Huh? . Yeah, I mean, I didn't really think about it. That would be a great application to, to, yeah, that would be wonderful. Great. Um, obviously you would have to proofread everything ChatGPT writes, because, you know, so, so I think the question is what controls you'll put in [00:45:00] place so that, um, so that it doesn't write a four chan post, but instead writes a summary of the patient's, uh, history.

Exactly. Exactly. Yeah. Okay. You can think big for this question. Okay. So Lily, if you weren't a product manager, what job would you be doing? It could be anything. If I weren't a product manager, I would love to go back to academia and be a professor. I think, um, I think it, it's a, it's also a very broadly, you could go pretty broad in that role as well and try a bunch of.

Fantastic. Um, I think department chairs, department chairs everywhere are, are chomping at the bit, and I'm sure you'll get a lot of emails after that response. Um, so another thing that we've seen recently are AI systems that can generate high resolution images of essentially anything that you want. So you can type a sentence like a cat sitting on a lawn chair, uh, drinking a Bud Light, and it will generate that for you.

Um, the question for you though is do you think things created by [00:46:00] AI could ever be considered art? Oh yes, absolutely. Right. I mean, it's already art. I mean, art is in the eye of the beholder. People say that. Right? So that, that is a definition of art. Yeah. Yes. And so if you think it's art and I think it's art, then it's art.

And I think it's art. So, and it is art. Oh, here's another alternative career move given, um, that you just triggered this memory about, , images being, being able to create a, a children's book. illustrator. That would be a, a fairly easy job. I, I feel like, I feel like that's an industry that will be disrupted in the next five years, given the generative language and art capabilities that Ai, AI has, that seems like something that you could bang out in like five minutes, uh, given access to those tools.

See? There you go. Side hustle. All right, will, okay. Will AI and medicine be driven more by computer scientists or by clinicians? Clinicians, [00:47:00] because they're the ones that have to use it and they're, they're the users. Uh, so it's in every field, you know, it's always the users that tend to drive adoption.

All right. And the last question of the lightning ground. Um, if you could have dinner with one person, alive or dead, who would it be? A dinner with a person alive or dead? This is a really tough question. So I'm gonna use the same. person I selected from high school in my high school college applications.

It was Marie Curie. Nice. Um, because I thought it was, it must be extremely fascinating to have a career like that in a time where she was, and also I believe she raised a daughter who was also a Nobel Prize winner. So that must have been a really, really fascinating. an excellent choice. So with the time that we have left, we're kind of gonna zoom out a little bit and focus on some big [00:48:00] picture stuff.

Mm-hmm, earlier in the discussion you mentioned that this really impactful paper that developed an AI system for detecting diabetic retinopathy had no home. It wasn't a computer science paper, it wasn't a medicine paper. It ended up in sort of one of the top medical journals, but still it was kind of like in between two lands.

And I find this cultural difference between computer science and medicine to be constantly fascinating. And I also feel that tension computer science. as a research industry tends to move very quickly. You publish in conferences. The conferences happen three or four times a year. There's a heavy preprint culture, so you're getting the latest science sort of on a continuous basis.

Medicine tends to be more conservative, slower, I think be more intentional in the types of things that it does. And so the question to you, and what I'd love to hear your thoughts on are what parts of medicine should emulate the computer science culture and vice versa. So what can computer science learn from medicine?

Ah, so I think [00:49:00] what I love about computer science is the way that we look at data. So data is there for learning, right? It's not there for, inspection. And so what I mean in that, a lot of times in medicine we collect data so that we can inspect and make sure that the thing that already happened went correctly.

And you know, if not, then corrective actions need to be taken. You know, um, it's, it just, it's really for either billing or sort of regulatory compliance kind of things, less so for learning. And what's interesting in computer science, or at least in tech, is that, every user that goes through the system really improves or try, or we try to improve the next user that goes through that system, their experience, right?

And so that, that's not, as big of a thing. I think in medicine currently, in many ways, we kind of don't have the data capture systems to measure and make that a reality. [00:50:00] There's a ton of interest in doing so, though. So I think that, given that we should be trying to harness all that enthusiasm, and move in that direction.

And then I think there's, in computer science, a lot of times we look at a problem from the, the hammer's point of view, right? The whole, you know, if you have a hammer, then everything's a nail. And we think, hey, you know, technology is going to solve all the problems. And so let's just train a better model, more accurate model.

And in medicine, you know, it's not just about getting the right diagnoses, right? It's hanging out in that area, um, of ambiguity and still trying to make really good decisions on behalf of someone else. And I think that's really tricky because you're also taking into account not just what a person's phenotype is, right?

What, what disease they may have, but also well, what goals they want in life, right? What is a good life for them? [00:51:00] And that's different for different. People. And so the ability to take into account a user's values in that way and also kind of hang out with them through ambiguity and through tough times, I think is really, really awesome.

And one of the best parts I think about medicine is to help people in that space. So, you know, you can see with Chat GPT and some of the other things like we don't, in computer science, we don't seem to understand ambiguity super well. Right. Uh, and so, you know, when we ask ChatGPT A question, it feels like there's no humility because there's just like, well, here's the answer, right?

I think we'll get there. I think at some point we'll program some level of, okay, we don't really know the answer, we're not very confident about this and help users understand the ambiguity and or the lack of confidence in the answer. But yeah, I think, I think those are the two pieces, that if we got.

both got the two working more [00:52:00] closely together, we'll see really interesting things for medicine and for, computer science. Awesome. So dealing with uncertainty and understanding values, um, mm-hmm. , that's fantastic. So, Lily, we've asked you a lot about your papers, both about diabetic retinopathy, other medical machine learning applications, but we wanted to ask you about a medical AI paper that you have not written, but in particular, what medical AI paper that you have not written has been most impactful on the way that, that you think shortly?

Well, I would say about a year or so after the JAMA paper, there was another paper that was led by a group of researchers. The lead author is Dan Ting from Singapore. But there were like a consortium of other, screening programs and scientists that worked together to essentially, they built a, I would say even a better model than our JAMA model.

It was around eye disease and diabetic retinopathy, some glaucoma predictions and whatnot. [00:53:00] But what was most impressive about this other paper also a JAMA paper, was a validation of the technology across all of these different populations. And I think that was a moment for me where I was like, oh my God, this is really real.

Right. I think the JAMA paper we were very excited about, obviously, but the big question was one in which populations would this technology generalize in? Like, could it be everywhere? Right. What's the scale and impact there? And then the second piece was, is this something that only like computer scientists can do, right.

Researchers at Google with all the, money and compute and whatnot, could do this or could other people do it right. . And so when we saw that and we saw that it was led by essentially, a group in medicine, right? I mean that found colleagues across the board, but a academia medicine, they were able, they did this and they did this better.

I would say it was like a moment of like, oh, oh my gosh, this is real. Like everyone can [00:54:00] get in on this and everyone can do this. And I think, I kind of think back to the lightning round question you asked me, which is like, who's gonna drive the adoption of AI as it clinicians or computer scientists?

I think there's a new breed of clinicians that appreciate and understand computer science and so I think there might be a, you know, they're clinicians by heart, right? Cuz it trained, but they understand computer science and data to an extent that they're natively bilingual. And honestly, those are probably going to be the people that catalyze change within the um, healthcare system.

Awesome. And that response is great for two reasons. One, it's very well stated, and two, I was gonna ask you what early career clinicians need to know about AI. But I think you just did that. So now I can ask you about explainability , cuz I'd like to come back to that. I have thought about explainability a little bit and usually it's in response to a clinician coming up to me and asking me about trust.

And so they will be like, how can I [00:55:00] trust this AI system if it can't explain to me how it works or how it's arriving at a decision? And I understand that because I also think that's part of the grooming process for clinicians that they have to round on their patients. They have to give like a systems level description of what's going on with their patient, why they're treating them this way.

And I guess the thing that makes me slightly uncomfortable is the mismatch between what goes under explainable ai. and what human clinicians expect. So as you mentioned, an AI can light up part of the region that it finds important for making a prediction, but it can't back out the type of causal explanation, that I think humans often want.

And so that's where I get off the explainability train where it's great for debugging. And so it's great for finding obvious model errors. This AI system is looking at chest x-rays and it's focusing on the hospital watermark and it's not looking at the lungs. Like those types of debugging exercises are very good.

Um, but I'm curious sort of if you think it's an integral tool [00:56:00] for bedside decision making or if it's sort of further up the pipeline than that. I think so. If I had to put my neck down, I think it is a little bit further up. than that. Um, I think the trust has to already be there when you're at the bedside, right?

So I, I'll give an example that's not ai, but you know, think about the treatments that we give people, right? We think we understand the mechanism of action, but. In reality, we don't really know the exact mechanism action, and we trust it because it's been well validated. That's why we, because there's an R C T behind the medication.

It's safe and effective according to the R C T. The MOA is kind of a story that we tell ourselves. Yes. Uh, the M MOA is a story we tell ourselves to predict side effects and predict what could go wrong. Right. And so the MOA is very important in part of this, to your point, troubleshooting and understanding the risks involved and, what kind of side effects we might see.

So in the same way, you can ask any clinician, and [00:57:00] you could, me included, is that like, can you precisely tell me how TNF alpha inhibitors work with my rheumatoid RA patients? Like, I probably couldn't tell you, but, um, on the spot, but there was enough trust built, you know, back somewhere in the back of my head where I've seen that R C T or like I've know about the R C T, I know that the approvals happen and that validation data is there.

So, And even if you haven't seen the R C T, you know that there's an approval process in place and the medication couldn't have gotten this far along without clearing several hurdles that you understand, even if you haven't seen it for this specific me, right? Yep, exactly. And I think what's actually really exciting about AI, and that's different than even drugs, lifesaving drugs like that, is that ai, the predictions are based on data.

And the data actually can sit in a monitoring system, right? A post-market monitoring system where, you know, right now with drugs, there's no data monitoring system, right? And so we rely on post-market [00:58:00] surveillance. So something goes bad, we report it, and then there's an investigation that happens. And the, and to close that loop, that takes a long time, right?

And so we are, obviously, we do the RCTs, we actually have trust that the approval happen, but sometimes, you know, like Vioxx or whatever, a bad thing happens, and then we find out after we'd say there's enough people that are harmed so that there's an investigation, right? . I think the advantage here with AI is that you could potentially see hotspots before too many people are harmed, right?

That, and so if we're able to actually put together systems, um, that collected in that way that do monitoring in near real time or just in time, then you have a continuous validation, not just a one point validation in time of a system. And you can look at drifts, you can look at different populations.

There's so much more you can do there. So I think explanation is a wonderful way of trying to predict what happens. And same thing as mechanism of [00:59:00] action, but where I would love to see more investment is the monitoring part in the continuous validation part. I a hundred percent agree with you there.

Yeah. So I think that's also a perfect transition to the next question, which is thinking about the deployment of AI across populations, do you think machine learning will exacerbate or will help mitigate or improve healthcare disparities? I think. The fact that we're having this conversation right now, and you're asking me this question means that it is more likely to help than harm, right?

I think many cases when health disparities, they're exacerbated, a lot of times it's cuz we haven't thought about it and then we're like, oh no, this thing happened. Oh, that's not what we wanted. Right? And so when we think about health disparities, we actually need to think about the design of the system, not just how the AI performs, right?

It goes back to like, who are your users? How are they using it? And if you are on the lookout for how this might actually exacerbate disparities, you actually put [01:00:00] in designs that mitigate that. Um, monitoring systems that actually look for it in the wild when you're deploying it, right? So you can't think about this as a afterthought.

It actually has to be part of the design and the fact that we're explicitly designing for this. Me have better confidence that this will improve in the future. I think in medicine, a lot of times people didn't set out to design a system that was unfair. Right? It happened because it was expedient to do X, y, and Z and then there was these consequences.

And so we really do need to correct for that by being more thoughtful on the design side as well as implementation side. Great. Um, now that we're over an hour having grilled you with tons of questions and your defenses are down, uh, I want to ask you to share what your most controversial opinion is, either in medicine, the intersection of medicine and ai, or just generally what's a deeply held, controversial opinion that you have?

Oh, man. . Well, okay. The facetious answer is like, don't you think [01:01:00] going back to, you know, my favorite fiction, uh, which is Harry Potter, like, don't you think Hermione should have gone out with Harry? What up with Ron guys? I don't know. I, I'm, I'm with you there. Like, uh, I don't know what the decision making behind that was, but the Hermione Ron love story always felt forced to me.

Yeah, exactly. So that was probably my most controversial opinion. Well, so I don't know if this is, you know, I don't, again, I don't know that this is a really a controversial opinion at this point. You kind of know what I think about it, which is the wonderful part of AI is really not the neural architecture or the fancy deep learning stuff itself.

It's the bread and butter stuff, right? Like how many times we get tripped up over image quality, right? You can't take a good picture. So, mm-hmm, you know, I don't know if this is a controversial opinion after all, or this entire interview has been controversial, but it's, it's really getting the basic stuff right that makes the biggest difference, I think, [01:02:00] um, in the world.

So, I think focus on the fundamentals, especially in an area like AI, can actually still be controversial because we tend to get distracted by the shiny objects and forget about things like study design and good data versus bad data. So I still think that, that there's plenty of controversy to be found in that opinion.

All right. We have our final question, and it's a positive one, which is, what are you most excited about in the next five years for medical AI? Okay, now it's triggering. Now you're triggering the whole other controversial opinion I have, and I think this is also what I'm most excited about.

So one of the controversial opinions I suppose I have, and also one of the most exciting pieces, I think is. how AI can make the participant's voice, the patient's voice louder and more central to healthcare systems decisions and whatnot. And so what I mean about this is specifically on the opinion side, is even [01:03:00] basic things like who owns the data, right?

We have this kind of really interesting thought about data ownership where it's like an either or, right? So either the patient owns it or the healthcare system owns it, but not both, that kind of stuff. And so I think the controversial opinion, the first piece is that I do believe that patients at the end of the day should own their data.

And I think many. people who are afraid of that actually there's a piece of like, how do we implement this? Right? And that doesn't mean that the healthcare system doesn't also own part of that data. It's not like, data's not like a piece of land. It's not like two people can't occupy the same space at the same time.

Right? It's data there, we can make copies of it. Right? Um, so there's just, the patient ultimately has probably the permission to share data and you know, which pieces of the record and whatnot. But it doesn't necessarily mean that healthcare systems and providers don't own pieces of that as well. So I actually believe that in current, a shared ownership model, but I also believe that there's a piece where [01:04:00] patients that actually should be, you know, be able to make decisions, give consent and whatnot about how their data is being used.

So that's one piece. But then I think the other piece about how AI and technology can help implement that is actually understanding what did I give permission to at what time and when I can revoke things. And so that's the technology part. But then, . If you also think about the way that we wanna engage with patients and talk to patients and say, Hey, you know, how are you doing right now?

When we wanna do that in medicine, we actually ask them to fill out a survey. Like a forum, like tell me how you're doing, plus or minus, right? But how people are doing that question is such a nuanced question. How you say it? What words you say all really matter. And so I think what would be really exciting is how AI actually can capture the nuances of how people are feeling.

Both, you know, in terms of like their functional abilities. Like, you know, can I walk down the street and pick up my mail without [01:05:00] being out of breath or whatever, right? But also like how they're feeling, um, what their mood is. Do they feel lonely and all that stuff, right? So usually if you wanna do characterization like that, you have to, like, it's a, a lot of manual labor.

It's really hard to do, um, hard to summarize, hard to capture that data. But with technology and with ai, like we can maybe do this, like scale this pretty easily, um, and then, and then be able to measure it, right? Um, measure how you're doing from day to day. And then have people develop treatments, interventions, uh, care management plans that then can move the needle to things that truly matter to patients.

I think one of the struggles that we have is we can't quantify, we couldn't, we can't measure the patient experience. And because we can't measure it, we can't optimize for it. So I think if we can start thinking about how to measure that, how to break that down into actual things that matter, I think we're gonna be able to find just new ways of helping people that we [01:06:00] hadn't thought about before.

Awesome. Before we let you go, um, were there any topics that you were hoping we talked about that didn't come up that you'd like to discuss before we let you get outta here? I'm not falling for that trick. Okay. Um, well, Lily, thanks for joining us. Um, again, I'm, I'm constantly impressed by your body of work, both on the innovation side and the implementations side. Um, and it's been a real pleasure having you today on AI Grand Rounds, so thank you. Thanks again. Oh, thanks for having me. And thanks for the fun questions. I could tell you guys, I spent a lot of time thinking through how to do this and how to make it, you know, not just a monologue. Thanks, we appreciate that.

Thanks so much, Lily. This was great. Awesome. See you guys later. Bye.