NEJM AI Grand Rounds Podcast TRANSCRIPT

Episode 2: Dr. Pranav Rajpurkar on AI and Radiology

[00:00:00] Welcome to another episode of NEJM AI Grand Rounds. Today we're delighted to bring you our conversation with Pranav Rajpurkar, who is an assistant professor in the Department of Biomedical Informatics at Harvard Medical School. And Pranav has really been a leader at the forefront of medical AI for several years.

Andy, I was really struck in our conversation with Pranav by his stories about things that can go right with medical AI, but also what can go wrong with medical AI. So aware it can learn to exploit parts of the way the problem is set up and solve a problem that you are not actually intending to solve. Yeah what I found really interesting about Pranav, having followed his work for a while is just how impressive his body of work is. What we touched on a little bit in the conversation is his earlier work in natural language processing, working with renowned computer scientist, Andrew Ng, and all the mentorship lessons that he learned from Andrew, and then how he picked

[00:01:00] medicine and specifically radiology to be his area of focus for machine learning and AI. He's gone on to make some seminal contributions in a famous model called CheXNet. And we learned all about that in this episode. And then we sort of get a peek into what Pranav is thinking about in the era of self supervised learning, which is such a hot area in AI right now.

So I really enjoyed this conversation Pranav. Was wonderful to talk to. It's a researcher that I thought I knew a lot about already, but I learned a tremendous amount from Pranav, during this conversation. Yeah, I think the, the circumstances around some of his big accomplishments in medical AI and the stories behind those were a little surprising to me, and I, I think it's, it was a really, really, really fun conversation.

So delighted to bring you all our conversation with Pranav Rajpurkar.

All right, Pranav. Welcome to AI Grand Rounds. We're excited to have you on the podcast. Thanks for having me. It's a pleasure. Pranav, a question we ask all of our guests, tell us about the trading procedure for your own neural net. How did you get interested in AI? What data and experiences led you to where you are today?

So for me, I would say [00:02:00] it started with a love for coding. I must have been in high school and I came to Stanford for a high school summer college program, and there I took the intro to programming course. I just thought it was so fascinating to be able to work on these problems, come up with solutions. So I went back, completed the senior year of high school, did more coding.

Then I really enjoyed building websites, and that's how I got into computer science as a field. And then my foray into artificial intelligence began in the first week of my freshman year. There's this week where we get to see lectures by people in different fields, and one of those talks was by Professor Andrew Ng, and he talked about helicopters that could fly upside down, robots that could walk upstairs, and I just thought it was just so fascinating and so fun.

So at the end of that quarter, I sent him an email. [00:03:00] I said, Hey, I really loved your presentation. I want to do AI work in the future. I can build web applications right now, would there be an opportunity to get into AI with you in the future? And he said, yeah, why don't you just apply to my lab?

And so I did and they liked my application. And so freshman winter, I was in a machine learning lab with very little knowledge of machine learning, but good amount of JavaScript and web development experience under my belt. Cool. Obviously you were interested in computer science, you were building web apps.

How did you get turned onto medicine? So what was your entry point for thinking about artificial intelligence problems in medicine? So when I started the PhD program at Stanford in computer science, I knew that I wanted to be working on the intersection of artificial intelligence and human intelligence.

I was very interested in that intersection. I didn't exactly know what in that [00:04:00] intersection I would want to work on. So I tried out a bunch of different things. I tried out security, tried out natural language processing. I tried out databases, social networks, and then in the second year of my PhD in the winter, so about halfway in, I came across a problem.

That was being worked on by someone in Professor Andrew Ng's lab that was looking at how we might be able to interpret electrocardiograms. These are measuring the electrical activity of the heart, and the problem was can we detect abnormal heart rhythms to the level of experts? And there was access to a large dataset collected from about 30,000 patients from this patch called I-rhythms Zeo-patch.

And I thought it was a fascinating problem formulation. Obviously, medicine is a very high social impact application of artificial intelligence, and I thought, let's give this a try. So tried it out, [00:05:00] and I remember that quarter really well. I used to go into the lab. I knew I had a milestone to hit in terms of performance, and I would try to continually improve the neural network that was designed to do this.

But even at the very start of that quarter, one of the first things I did at the encouragement of both my colleague and my advisor was read a textbook on EKG interpretation. So I remember going to the medical school library for the first time and going to the basement where there were these rolling shelves that you had to roll to expose the shelf that corresponded to cardiology.

Picked up the book. And worked my way through it. And that was great. It felt very useful to be able to immerse myself, understand what problem I was solving, and then be able to say, okay, now I understand it sufficiently well to at least know what I don't know, and sufficiently well to say, I can start building something for this.

And did you work with [00:06:00] clinicians on that project? So on that project I was more on the development side, given that we had defined the problem. And so there had been some legwork done. I would say that was the exception for me. Since then, I have had a lot of work that was very closely collaborating with clinicians.

Yeah, I think that many folks might be envious of that entry point into medicine. Cause as I'm sure you know, usually you get this very messy dataset, kind of an under defined or underspecified question, and then you kind of have to navigate this maze. So it sounds like that was a really great entry point for you where there was a hill to climb and you have a way of climbing that hill and it was just putting your head down and working really hard to do that.

Yeah, that's right. I would say I wasn't so lucky always in the future. Yeah. So I remember right after I finished that project, we put out a preprint and a lot of people really liked it. We basically got to the point at which we said we can perform as well as cardiologists in the interpretation of single lead [00:07:00] ECGs.

So that was great. And I remember right after that project, a colleague of mine and myself, we decided we would email every faculty we could find in the med school who had some interest in machine learning that we could find. So we made a list and we just sent out all these emails. And we got a few responses.

Those were, again, very lucky. One of them was a, uh, response by Curt Langlotz and Matthew Lungren, who would be, and still are my long-term collaborators, in which they said, we're very excited that you're interested in working on this. We have a list of problems, and the first one of which was bone x-ray interpretation and chest x-ray interpretation.

So we worked on that and that was a very good project as well. It was intellectually very stimulating. Right after that, I thought, wow, I've had two successes in a row. This is gonna be easy to try to take to other [00:08:00] problems. And so I remember working on this project, which was an ultrasound project, so it was new modality.

I had not seen it before. And we had this repository of code where the idea was you could plug and play. Right. So if you have a new modality, this was ultrasound and we could do the detection of DVT. So blood clots. And we had this pipeline and I was very proud of the pipeline being able to run anything in about 24, 48 hours.

So we thought, we have this data set, great, we got it, we have the labels associated with it. Why don't we just feed it through this toolkits? And so we did. And we got 0.95, AUC. And there I'm thinking another great success. Another one. This medicine stuff is easy. Well it was, it was just amazing. It's like we have this tool and it's working so well.

And this was, now we're talking January, 2018. So this was still early. We hadn't seen so many successes of deep [00:09:00] learning across all of these different modalities. This would be a big deal. And this is when we thought, okay, let's take a few verification passes through this. So code looks good. Let's look at where the model is looking.

And so we look at where the model is looking and we have a classic ultrasound image and it's looking at the top right corner. Ah, I thought, okay, top right corner. What's that? And it turns out that's where your machine metadata is, is you look at the top right corner, you see the machine metadata and that can make all the decisions.

It's like, wow, that doesn't sound right. And it turns out there were two kinds of machines in where our data came from. one of them was a screening setting where there was almost no clot. And the other one where there was always clots. Mm-hmm, it's like, oh my God, this is just, you know, this is problematic, but we can fix it.

We can fix it by just cropping that out. So we did that, cropped that out, and I knew we weren't gonna do as well back to [00:10:00] 0.95 AUC. And I thought, okay. You know, this time it's gotta work. And so this time we looked at the image and it was pointing to where you would expect the clot to be, but not exactly in the center.

It's sort of to the right or to the left. And we saw there were these xs near the clots. Why are there xs near clots? So then we thought, okay, we're doing this all wrong. Let's go and talk to the people who are collecting the data. Cuz we'd been talking to radiologists. This was multiple clinicians and engineers involved.

So we did that and we go to the hospital to speak to the stenographers. The sonographers say, Hey look, you know, when we find a clot, we mark it, we mark it with these two Xs, we're we're measuring it. And then it struck me if they're marking. The task is to find the X in the image. It's not to define. It sparks the spot.

Right, exactly. And so that was crazy. And [00:11:00] so then I thought, okay, I asked them, Hey, so what do radiologists do on this one? No, they don't do anything. We do the work here and look, I can make a clot look not like a clot, and I can make not a clot look like a clock. Look at the way I'm doing this. And that, for the first time, and then this was a lesson I would carry through the rest of my time, is it's very important to understand the data generating process and a lot of these ideas or hopes that we can have plug and play in terms of the technology is a step in the right direction, but certainly not at the expense of everything that has to happen in terms of the task definition.

Mm-hmm. and the understanding of the data generating process to speak to your point about needing to go about understanding the data set and its noise. Yeah. I always think that that is one of the fundamental challenges of machine learning and medicine is aligning the data collection process or the data generation process with the task that you actually want the algorithm to learn.

Cause it's gonna learn something, but it may not be the thing that you [00:12:00] actually wanted to learn unless that alignment is very tight. So I think that again, it's great that you learned that early on and obviously it's carried with you through the rest of your project, which I think have been very good examples of getting the alignment part correct.

So it seems another theme from that early work in medical machine learning is that it's critical to understand the data generating process, especially because the models will ruthlessly exploit whatever. Pieces of information it can exploit, which out of sample or when the model is applied in the real world, might not be there or might be totally different in another population.

I'm curious where this ranks for you, Pranav on the sort of list of things that keep you up at night when you're thinking about these models being deployed out in the wild. Is this the top problem? Is it amongst the top few problems? Or is it potentially something that's over-emphasized as a concern in the field?

Sure. So maybe I should take a step back and tell you about the way I think of my labs work. So I would say the one line pitch is innovation to make medical AI intelligent, safe, and [00:13:00] useful. I think all of these are very different directions. So intelligent is about how we can push the capability of these systems to do more and more of what we would consider intelligent medical decision making behavior, whether that's I can make the right diagnosis for this patient, or I can make the right prediction of the future state of this patient, or I can make the right treatment decision recommendation for this patient.

The safe part relates to, I think now the most important question being will this model apply to different patient populations, to new geographies, new clinical settings? I think that's the big safety question of today and how do we help the community transparently measure advancements on that? And the third part is, useful and I think on useful, you don't necessarily need intelligent solutions to begin with.

In fact, if we try to understand [00:14:00] workflows and try to think about the ways things can be improved, a lot of that could be drawn to interfaces or the right timing of doing certain operations or the right attention. But these are obviously very necessary if we want to think about showing a relationship between having a decision aid from AI and saying this actually improves some clinical outcome that we care about.

And I think it's the mix of all of these different things that we need to be able to say in 10 or 20 years from now, medical AI is gonna significantly advance just for all of us the way clinical care is going to happen. That's great. So just to broaden out from medicine for a little bit , some of your early notable work includes SQUAD the Stanford Question Answering Data set.

Could you tell us about SQUAD? Yeah, so it was 2015 December that I was at NeurIPS in Montreal [00:15:00] and I remember there being a lot of talk about natural language processing being the next big frontier that AI was going to be able to tackle. Now, six years later, we know how that has evolved, that that prediction did end up coming very true, but at that time there was a lot of hope associated with that particular application.

Can I just get you to briefly summarize what folks mean when they say natural language processing? Sure. So, I would summarize it as how can you read and understand text to do useful things. Mm-hmm. and useful things could be how do you answer questions or how do you summarize something or how do you fill in a blank?

Mm-hmm. , given the context around you, these are kind of questions that people were interested in natural language processing and one of the big open problems at the time was when we Google search [00:16:00] something, if you Google search a question, it can come up with a list of pages that are relevant that contain the answer, but what the step that was missing was from those pages.

Can you extract exactly. The segment that corresponds to the answer of interest. This is the problem of answer selection, and this was a big open problem at the time. And I was going to start working for the next three months with Professor Percy Liang. This was a part of a rotation where for three months you work on a certain topic with a certain advisor at the start of a PhD.

And so that was, that's mine. And we spoke about, well, one, that being a great problem, and especially for me because I came from a web development background and it seemed natural to be able to say, Hey, how can we apply some of your web development skills onto this? And so the idea was could we generate a large data set.

Of the pattern that you have, passages of [00:17:00] text, you have questions on that passage, and then you have answers which are segments of that passage. So for example, you might have a Wikipedia passage about some topic. Someone would ask a question and also provide the answer highlighting the span of text that answers the question.

And the largest data sets for that were relatively small. And, we said, okay, we can make that bigger. And so we ended up with SQUAD, which was a hundred thousand questions for reading comprehension of text. And, it, ultimately became one of the most cited papers and natural language processing of the, of the last five years, counting from 2016.

Not bad for a rotation project. Very good for a rotation project. That's right. A lot of that I would attribute to very good insight into what was a problem worth working on. And then, The data set itself a sweet spot of having a data set that was not too easy and not too hard, that people [00:18:00] couldn't make any progress or people would make progress very soon.

And so I was responsible for the squad leaderboard where people submitted their models and I would see the results that they got. And we saw the rise of attention models. This is this new class that is so prevalent now that showed on the squad leader board they could achieve the first rank. I remember one night, this must have been 2017 or so, I got an email from someone at Google and they said, Hey, we have just uploaded a model.

Would be great if you could help us test it. We have a deadline coming up. And so I was like, okay, gotta test it. And I couldn't believe the numbers that they got. So they were not only first on the leaderboard, it was first by quite a large margin. And this model they called Bert. Mm-hmm. And so Bert became a household name in machine learning, which was a very interesting novel way of being able to [00:19:00] pre-train a model and then apply it to useful tasks.

And it was amazing to watch that development, those breakthroughs unfold at a ever accelerating pace through those last few years. And so that was a great lesson for me in working on the right problem at the right time. I'm sure that a lot of the lessons that you learned in creating a benchmark task must inform what you're doing in medicine because in many ways, we don't have great benchmark tasks for a lot of medical tasks, and so that you must carry some of those lessons forward with you even today.

I think it's important to have alignment on two things. One of them is, do you and I agree on what are important problems to solve? And then do you and I agree on for that important problem, do we have the right data set to be able to answer that important problem? And I think without alignment on these two people will

build machine learning algorithms for [00:20:00] things that are not going to ultimately be helpful or translate into practice. And I think this model that the people who understand the clinical questions and who have access to the data, one of the responsibilities here is can we define that question clearly and make the data available to answer the question clearly available to the machine learning community so that people who are good at developing these models can actually work on good problems and the lessons that we learn there

we can directly apply into models that translate. I think completing that loop is one thing that I hope that we as a community, which consists of many different kinds of skill sets and people with access to different pieces of the puzzle can get right. Got it. So a few more questions about your background.

So you've mentioned some of your collaborators who have informed you, you kind of touched on your PhD advisor, Andrew Ng, who, as most people know, is this sort of very big name, a luminary in the field of machine learning and artificial intelligence. Could you tell us a [00:21:00] little bit of what it was like to work with Andrew?

What his management style is like and sort of, I always think of my PhD advisor in substance as being my parent and I carry a lot of lessons from them, with me today. So could you talk about that a little bit? I learned a lot from. I'll give you a couple of examples. So one of them was the way he thought about impact.

I remember I was working on a project which would help Stanford students. I was a Stanford undergrad at the time, navigate through their academic journey. So help them find the right classes as very passionate about this. It was my pet project. It also kept me up to date with the latest web development technologies.

So I had a lot of fun with it. I was chatting with Andrew one day and one of the things you said related to how many people are you in effect. and at Baidu used to think about HMU's, which was hundreds of millions of users, and that [00:22:00] was a benchmark by which you could decide whether or not to work on a problem.

Is this going to affect at at least that many users? And that's a very interesting way to think. It's to say, I'm going to select problems based on a impact that I'm defining, and if it doesn't meet that threshold, it's not a great problem to work on. And so I took that away from him. And it partly, that's what has defined my choice of problems to work on, which is, one, I worked on X-rays.

So X-rays is the most common imaging modality in the world. 2 billion chest x-rays taken per year. Another modality I worked on was ECG. There are 200 million ECGs done annually just in the US and these are modalities that affect a lot of people that are worth tackling first, before we go on to modalities that affect fewer people, was the philosophy I took away. In terms of teaching.

There's my second example. As you know, Andrew Ng [00:23:00] is a fantastic teacher. Millions of people have taken his courses, and one of the things that I learned when I started teaching, especially in the online setting, was this idea that people shouldn't feel like they're not sufficiently ready to be able to understand this piece of the puzzle.

So really being able to distill what's important to understand and what's really not important to understand, and I think this idea of how you give psychological safety or even that you give psychological safety. Mm-hmm. is one thing I found useful to take away in my philosophy as a teacher, you have to give students the attention weights.

Yes, that's right. You know what to study and what to ignore. Exactly. Yeah. Yeah. I wanted to follow up on your first point there about impact as a researcher. One of the things I struggle with is short-term, especially as junior faculty, which we all are, is short-term impact in doing the thing that can affect the [00:24:00] most people.

Now, balancing that with long-term research interests where maybe there's no feedback for that today or no impact today, but I'm working on something that has a longer time horizon. How do you balance that in your own research portfolio? As for, Okay. X-ray is the most important imaging modality today, but I know that in 10 years we might be able to do something else.

How do you sort of think about that balance? Well, I think about the end goal that I want to have and then what I would consider as victory milestones towards that end goal. And if those victory milestones can be done in periods of six to nine months, I consider that to be a good milestone to have.

That's neither to dissociated with the final goal, nor to advance for there not to be measurable progress in between. And so I've found that to be the sweet spot of projects that take between six to nine months to complete. So figure out where you wanna go and work backwards from that and break it up into sort of bite-size milestones that, you know, are [00:25:00] breadcrumbs on the path towards that bigger agenda.

Yeah. Cool. So Pranav, one last background question before we jump deeper into your research. What medical AI paper that you didn't write has been most impactful on the way you think? I think for me, I'm going to answer that question two ways. One of them is, for me it was a book, not so much a paper, which was the Creative Destruction of Medicine.

Hmm. So this was when I was starting to get into medicine, selecting that as the area of interest where I wanted to apply my AI expertise. I read this book, it's written by Dr. Eric Topol, and it was great in a couple of aspects. One of them is it was very clear that there were opportunities for improvement in the field.

Mind you, this was at a time at which I knew very little about medicine and how much there was in terms of improvement [00:26:00] in terms of being able to do, sensing of different health aspects, using, ECG signals using other kinds of remote monitoring and how much. Opportunity there was to say the future ultrasound.

Being the new stethoscope was one of the ideas that I learned about through that book, and that seemed very interesting as well. And I remember after finishing that book thinking, wow, if I ever think that we're out of problems to solve, I should come back to this book and think about just how much has not been done that we can very clearly see and lay out will likely happen in the future.

And so I think that was an inspiration for me. I would say the second class of papers, I'd say for me is not so much a medical AI paper. I would say an AI paper in general, is that I've really enjoyed some of the works [00:27:00] on large language models that have come out. The kind of paper that I enjoy is one where I think this is very elegant.

and this is always the way things should have been done and extra bonus points to me. If I had already predicted that I would've read such a paper in the future, I think that makes me feel good about thinking about, okay, I have the right idea of where the field is headed. I think some of the works, including the contrastive image language, pre-training work from open AI we're very much in that direction of saying, I can learn from data that's naturally available in the form of images, in the form of texts, and have my learning allow us to communicate in language when we're given these images.

Well, Pranav, we talked a little bit about your work on what is arguably, the world's most impactful ML rotation project in SQUAD . Uh, so now I wanna take you back to 20 16, 20 17, and I remember this moment [00:28:00] because there were really three papers that came out that convinced me the age of medical AI had arrived.

And it was a paper on detecting diabetic retinopathy published in JAMA. There is a paper from Stanford on diagnosing skin cancer using deep learning. And then there was your paper on chest x-ray interpretation using deep learning. And the name of the paper was called CheXpert. And so I was hoping that you could give us a little context, a little more background.

You sort of teased us already with a little bit of the origin story for CheXpert. What was CheXpert? How did it come about? And how did the clinicians that you were working with at the time really inform that project, right? So I'm gonna have to walk two steps back from CheXpert to CheXnet and then build my way up.

So this was October, 2017. And this is a interesting story of how small groups can actually do non-small things. Mm-hmm. I remember it was a weekend and uh, the NIH had come out with a data set called Chest X-Ray 14. Mm-hmm. . It was a large [00:29:00] data set of about a hundred thousand images. And at the time I was starting to lead or co-lead a program called the AI for Healthcare Bootcamp, in which the idea was undergrads who had done machine learning classes would get a taste of what AI research looks like.

And I was in charge of coming up with a medical imaging project. And so the camp was about to start the week after and I thought, wow, this would be a great project. So I texted a colleague of mine, I said, Hey, this is probably a great project. Let's see if we can set this up, put together some starter code and people can work on this.

So I emailed my radiology colleagues as well. Hey, this data set has come out. What do you think? Got some confirmation that it was good, things seemed well defined in general. So I put together some starter code over the weekend. This program started, and uh, this was assignment number one. And as a [00:30:00] group of six people, all of them doing this assignment individual, I'm getting this stinking feeling where there's gonna be this super impactful paper that happens as part of like an undergrad project.

I feel like it's, we're gonna have SQUAD part two here. Yeah, yeah. No, that, that's exactly what happened. So just in a couple of days we had a leaderboard on our internal website, and there we had the results that the NIH chest X-Ray 14 paper had themselves written about, right? Which is, these are the networks we tried, this is the performance we achieve.

And we saw within two days we had already beaten that. And so we were already showing much higher performances on these different categories. And I remember thinking, wow, there's a large space for improvement and we've shown a lot of improvement. Let's write about it. And we started writing about it, and this idea came up.

Oh, you know what? We should measure what happens in terms of how much room for improvement is there compared to a radiologist. [00:31:00] So let's pick one task. Let's do pneumonia detection. We know pneumonia is something that affects a lot of people worldwide, and a diagnosis can happen, or detection of evidence can happen off of chest x-rays.

So let's do that. And so what we did was we took a subset of the test set. We said, let's get it labeled by, I think it was four people or so, four radiologists at Stanford. And then we said, how will our model compare? And I remember thinking, this is crazy because we saw that if I didn't tell you which one was the model, which one was the radiologist, they would look very similar to you.

And so the data was basically saying, okay, there was evidence that you were able to achieve. Radiologist level detection for pneumonia. So the way to approach a machine learning paper writing is you write the paper you uploaded on a Preprint server, [00:32:00] which is archive, and then later you submit it to a journal.

So we uploaded it to the Preprint server, and I remember us doing social media. So my co-advisor, professor Andrew Ng, tweeted, uh, the paper out with the, with the question, will AI replace radiologists? I actually have the quote here. Can I read it? Yes. Uh, should radiologists be worried about their jobs breaking news?

We can now diagnose pneumonia from chest x-rays better than radiologists. And it was a great question. It's a very fair question and it got a lot of people excited. And then it got a lot of people debating, was that an even reasonable claim to make? Did the data support the reasonableness of that claim?

And it was very interesting from my perspective because it was very much we got thrown into the spotlight. Mm-hmm, and now we have this claim to back where it's clear, you know, we did this with four people, one task. And so we immediately started working on, let's [00:33:00] scale this up, let's scale this up to all 14 pathologies in that data set and let's scale it up in terms of number of radiologists.

And so this now took time and then we finally published this in Plos Medicine November of 2018, where we showed to anyone who was saying the data wasn't enough last time. And to ourselves that across all of these different pathologies you could do just as well as radiologists would be indistinguishable.

And that was the, the big result of that paper that was called CheXNeXt. Uh, next being sort of the the next, yeah, there's a whole CheX cinematic universe now, and we're gonna touch on some of them later too. But there's a whole bunch of CheX, variants now. Exactly. Exactly. But just at the time this was going on over the past year, we had had a lot of learnings and the field had a lot of learnings.

One of the learnings was if you looked at the NIHS data set, people disagreed with a lot of the labels. There was a lot of noise in the [00:34:00] labels. And this was very famously explained in a blog post at the time by Dr. Oakden-Rayner. And the main finding there was, we cannot trust these labels for evaluation.

But the key here in my mind was, we can trust these labels for evaluation. Not for training. We can still do very good training with these labels. It's just that we can trust the ground truth when we come to evaluate it. How can we solve this? Well, assuming no other context is necessary. If the image is enough to make a detection or a diagnosis, then why don't we have several radiologists read the same image and let's take their majority vote?

And that's what we did for this study. But it did bring to light this idea that we really need clean data sets that have labels that we can generally trust. And that was the motivation behind us working on CheXpert, where we had a discussion. So Matt [00:35:00] Lungren, Curt Langlotz was involved here, and we said, okay, we have data from Stanford.

Can we do. a clean labeling of that, improve on the natural language processing there and be able to make that data set available to the world. And by the way, let's replicate the CheXneXt study in which we're again able to show that we're at least close to the performance of radiologists. And so that was the CheXpert effort.

And we published this in Jan 2019 in Triple AI, where we said, here's a data set. It's been labeled cleanly. Here's again, data showing we can do really well. But by the way, let's also have a competition around this. So let's track people making progress on this task of chest x-ray interpretation to a point at which we can be clear on this data set.

We have surpassed the Stanford radiologist who read it in terms of the performance of the models. [00:36:00] One of the nice things we were able to do at the time is there was a group at BI-MIT- Harvard at the time that was also parallelly, creating a data set of chest x-rays. And we said, Hey, why don't you try out our labeler?

And so they tried out our labeler and it worked really well, much better than what the NIH's Neg Bio Labeler was doing at the time. And so both data sets, which came out nearly at the same time, collectively having about 400,000 or 500,000 images had the same labeler that was applied to both. And that to me was a win in terms of what we were able to achieve in terms of how well something can transfer to another hospital, but also a joint effort between two different groups to be able to work towards a shared goal that's, that was gonna help the community for the years to come.

So it seems like there's both with the world's most impactful rotation project, machine learning [00:37:00] rotation project, and with the early CheX work that led into CheXpert and then CheXpert itself, that.

I'm trying to isolate the common ingredients of success, right? And I'm first struck by how non-linear progress can be. And so it can be very, very, very fast. At certain periods where you recognize an opportunity, you quickly disseminate something into the world, get some feedback, get some criticism, which is useful to then guide the next round of bigger projects ideally a preprint actually leads to. And so I think this is very common in the computer science and machine learning community. Unfortunately not that common in the medical community. In medicine. We typically like to wait until something is clean, polished, ready to go, and then publish it. So I'm struck by that.

I'm also struck by the way in which you're able to release data sets at this scale and very grateful as a researcher, and I think many of our listeners are as well. Another ingredient seems to be leaderboards, so you seem to be making this into [00:38:00] competitions.

And then very quickly getting a pulse of where the best performing models are and where the most active currents in the field are to then jump on , and to build off of. Another ingredient here I think is also very strong collaborations that you've built with clinicians and with radiologists specifically for CheX.

So I'm curious what you see as the sort of the key ingredients for building success in medical machine learning applications. And also, you know, how you build those collaborations with other clinicians, with radiologists, with cardiologists in the context of putting out these massive data sets.

Yeah, I don't want to downplay the impact of having very good teams that work closely to get these problems solved. And the structure for a lot of these projects, including CheXneXt and CheXpert, and a lot of this work has actually been very driven by fast moving. Young researchers, often undergrads or master students that I was working [00:39:00] with at the time in collaboration with often senior colleagues in medicine or in machine learning.

And it's really been quite crazy to see how small interdisciplinary teams that are well-structured in terms of how there is a good focus on the right engineering principles can actually make, in terms of having high impact. I think in terms of finding collaborations, I think there is a large gap on both sides of people who are engineers who are interested in applying their skillset to something useful and people in medicine who face problems all the time, that they would like someone to solve for them. And I think this great matching problem has for me, had a simple solution of I will [00:40:00] try to email as many people as possible.

Love it, to, to try to find these collaborations. So let me tell you about a, a project that I'm working on now. It's called Medical AI Data for All. And my hope with this project is to create the largest in terms of number of hospitals involved, dataset for starting with chest x-rays, but expanding to all kinds of medical images across the world.

By having a small amount of x-rays that are shared by every hospital publicly, but standardized by us, such that you can now evaluate whether models are going to fail or succeed when you apply them to different places. It's like , a test bed of sorts. Yeah. This is a test bed, and I think this is the greatest challenge of our time in terms of safety evaluation, which is simply that we have no data [00:41:00] except for anecdotally what we're hearing from medical AI companies that are bringing their solutions to new hospitals and finding that they have to retrain their models to make them work on new hospitals.

We have no public way of having the community transparently measure whether or not there are pitfalls in the way that we're saying models will generalize from site A to site B, but now let's approach it in a principled way. Well, to do that we need data and we need data of the right kind. And this effort called medical AI Data for all has had my team reach out to hospitals all around the world, establish collaborations and say, Hey, here's our pipeline.

Here's our ask, and here's the way that we're gonna go about this, collaboration and the science are you in? And a lot of people have said Yes, and I think this is how step by step we start building communities around how the future will work. . [00:42:00] Great. Maybe to just ask you about some of the more recent extensions that built off of CheXpert and the CheX multiverse.

We saw your recent paper in Nature Biomedical Engineering. I'm gonna read the title because there's a very key word in the title I think that I'd love for you to talk more about. So the title of this paper is Expert Level Detection of Pathologies from Un Annotated Chest X-Ray Images via Self Supervised Learning.

So un annotated is the striking word there to me. Could you give us an overview of this paper and CheXzero? What was your main question? What did you find? Sure. So I think this started as a dream several years ago where I have to contrast what we did here with how things were typically done.

So let's take CheXpert, right? We've talked about CheXpert. So with CheXpert, we built this labeler that was able to say, given this radiology report, let me identify all of the diseases that were mentioned in that report, and let me extract that and [00:43:00] that I can use to train a model that will say, in this image, there are these particular diseases like plural, effusion, pneumonia, cardiomegaly.

And to do that, you have to spend months developing this labeler. , it has to be a person who is in charge of engineering. How you find the different terms. And so we have this table that says, how is endotracheal tube referred to Uhhuh in text? Well, there's ET t, there's endotracheal tube, there might be misspellings.

And so we have to engineer this over months and months. Medical notes are famous for having their own vocabulary abbreviation, sort of non-standard English. So it's a non-trivial thing to extract those key terms. Exactly. And to me this seemed like a big problem because now let's say you say, oh, I don't want 14 pathologies, I want 28.

I need to go back and develop for several more months to say, okay, now I can detect 28 pathologies. Cause it took me so much [00:44:00] time. The other alternative, which a lot of resource rich companies and institutions can take is to say, let me just get a person to read through all of these chest x-rays and annotate explicitly pneumonia.

Yes. Plural effusion no. Normal, no. And you can do this at scale, and then you can calculate the cost of this. It's very high. It's high in terms of the time. It's high in terms of dollars if radiologists are being paid to do this. And so the dream was, well, could I build a algorithm that didn't need all of this explicit training data, but instead could learn by reading the radiology report that's available.

And that's how it learns to detect individual diseases. And so that's what we showed in Cze Zero, where we were able to say, we're gonna look at pairs of images and reports, and we're just gonna learn to identify does [00:45:00] this report match this image? And you say, if the model says Yes, this report matches this image and it's right, we credit it and otherwise we penalize it.

And so that's what the model starts with. And then we can apply this very neat trick where we now have this train model and to ask it whether or not something has plural effusion. We create two reports. One of them we say there's plural effusion in this image. One of them we say there's no plural effusion, and we see which one of those two report texts better matches the image.

It's a relatively simple, straightforward idea, and it is built on two works that inspired this idea. One of them is Open AI's Clip Model, and the other one is Stanford's Convert Model, which was actually an original inspiration for Clip. But this is a elegant idea that we were happy to find, [00:46:00] actually came to the same level of performance as we had shown in the early CheXneXt work, which is to say, well, yeah, we're about the same as, as radiologists on this task and

of course there's still room for improvement. The best supervised models can take you even further. But this is a big proof concept because now we've been able to show, hey, we can take our development time to a whole new level because we've just cut down months of work that was required to be able to reach high levels of performance.

And so that to me is very powerful. And I think one area where we're gonna see a lot of advancements in the next few years, and maybe I can just interject here and gush a little bit, um, because in 2016 or 2017, I remember kind of being floored that kind of out of nowhere. There are now systems that, we can debate whether or not they're human level performance or subhuman.

They still fail in kind of embarrassing ways, but nonetheless, like overnight, these systems that can [00:47:00] interpret chest x-rays kind of suddenly appeared. Um, fast forward five years now we have systems where you don't even need the radiologist to actually annotate. You can give them a high level description of the radiograph and it, in some sense is reading the chest x-ray report to learn what features are present in the image or not.

And then when you wanna label or get an interpretation for a new chest x-ray, you're literally just asking it is pneumonia present in this chest x-ray? And it will give you a back, essentially, like a yes or no. So it just seems to me like there's been amazing progress over the last 5, 6, 7 years. And I find it very exciting that now we're in this sort of very, we don't need labels, but now we can interact with systems using natural language because you're literally asking the model questions about what's in the, the chest x-ray.

So again, I've been floored. I wonder if you. Think about what CheXzero 2.0 looks like. What, what new capabilities it might have. Yeah, so we've talked about reading radiology reports as a way to learn. I wanna get to writing radiology [00:48:00] reports in their full entirety. Mm-hmm. , I think one of the shortcomings with current systems is that we think of simplifications of the full clinical task in diagnostic radiology.

There are several simplifications that we make for chest xray interpretation. Let me share a couple. One of them is we think about defining certain diseases and we say we're gonna give labels for each one of them, predictions for each one of them by model. And people have scaled this up. I think I saw one with 127 findings.

But if you read a radiology report and you try to understand what is the information being conveyed, It's not just information about whether something is present or absent, there's attributes associated with that description. There's attributes associated in terms of location, in terms of severity, and in terms of certainty.

I think all of these [00:49:00] are well captured in language. Language is the form of communication that's most natural in terms of the current clinical workflow. But one task that has long escaped our machine learning formulation is saying, can we directly get to the radiology report from the image without actually going through this intermediate step of saying, I'm only going to make these selected predictions.

And the power of this is that we'll be able to have basically an input image and output a radiologist report to the point at which I hope it will be indistinguishable or even better in quality. to the physician that needs to use the report than the current reports are, and the use of this is in terms of standardization, in terms of saving time, but then also in terms of improving access in places where there [00:50:00] isn't going to be a radiologist maybe at night.

Then you can have these systems take that place, and of course there's a lot of regulations that need to change to make this future happen. But I think technologically, this is where we're headed and this is where I think there's going to be a new wave, a new generation of medical AI systems.

Do you see this as changing the role of clinicians in these medical machine learning projects? Do they get to spend more time on other aspects of it? Does it increase the importance of the traditional role of clinicians on medical machine learning projects? So, in my mind, maybe let me, let me describe to you the, the aspects in which I think clinical expertise is key.

I think number one is problem definition. I think in terms of problem definition, there is. Big difference between a problem definition that works 90% of [00:51:00] the time and a problem definition that works 99% of the time. And I think if we wanna say these systems are safe to use at all times, then we need to cover that gap.

And so I think if we think about a roadmap that needs to happen from where we currently are to what is the full complexity of the clinical task, there needs to be a lot of defining that needs to happen to get us through that roadmap. Maybe even change in guidelines, maybe change in terms of standardization of terminology.

There's a lot to be done there in terms of now saying, well, earlier I had systems, which were people who were at a single institution that had their own style, that had their own workflow, to now we have these systems that are gonna be across all of these different workflows in setups, which are gonna be very different from each other.

What is the standardization that we need to have in terms of the way we set up our tasks, [00:52:00] in terms of the way we set up our workflow that's gonna make this truly work? And I think that's a really key part of where clinical expertise is going to be required. And that's one of the primary ways in which I think clinicians are gonna be leading the way.

So just to switch gears a little bit from research, you've been very involved in medical AI education initiatives. We spoke briefly about your online course with Andrew Ng teaching AI from medicine as a specialization now on Coursera.

You are the director instructor for a new course at Harvard CS 1 97 AI research experiences. Could you tell us about your experience, both with the massive online course and then a traditional course at the university, whether you see big differences between the way in which you design those courses, or whether there are more similarities between those different efforts.

Yeah, that sounds good. I think one of the reasons I [00:53:00] decided to become a professor, Is that I really enjoyed teaching, and this has been a longtime interest of mine. In high school, I used to tutor econ in business, and then when I was doing the PhD, I felt like all the things I was learning at the intersection of AI and medicine were scattered in terms of where I could find resources about them to be able to catch myself up that in the third year or so, I thought, I wish this would be a course.

I would've learned so much if all of the things that I've learned would be structured in a way that I could just go through something for a few hours and know exactly what I need to know to solve a variety of different problem formulations. Finally, at the end of my fourth year, I decided it was time I should do this, and I wanted to do it as a Stanford course, and I remember talking to Andrew Ng and he said, [00:54:00] why don't you just do it on Coursera for the World Impact

That's right. Um, a Hundred Million Users. Yeah. Hmu gotta get those HMU's. So I thought, yeah, that sounds great. It would be a challenge for me to try to teach something that would be seen by a lot of people. And I remember spending a lot of time thinking about what are the skills someone would need to be able to understand any paper in medical AI.

This was a time at which I was trying to keep up with every single paper published in a major journal and trying to distill what are ideas that I have learned from other places that I can put into this course. So ultimately this ended up being three courses. So it was a specialization where part one is on medical diagnosis, part two is on medical prognosis, and the third one is on treatment and it combines basically toolkits that you would need for each of these different problem [00:55:00] formulations and tries to explain them. And I would say my estimate at some point was, for one minute of video, I was doing 24 hours of preparation. Wow. And it was very hard, but I felt like the clarity of thought I got by finally distilling it down to a minute was just very, very high.

And I really enjoyed that and pretty proud of the specialization we were able to build. And I hope that has been a useful resource for the community. And I hope it'll continue to be a useful resource for the community, even when the technologies that we're using are changing so fast. And correct me if I'm wrong, it's primarily targeted at machine learning CS types.

Do you have any plans for a clinician side of that? Yeah, I hope to continue teaching in the future to different audiences. I think of my own pathway as someone who was [00:56:00] trained in machine learning and then got into machine learning for medicine as a pathway I understand well, and so that's why this was targeted towards people who came out of maybe a, a machine learning class, which is really well taught on Coursera and then now wanted to contribute their skills to problem in medicine.

And then you asked about how this contrasts with my recent class. So CS 1 97, for which the notes are all public by the way, so you can follow it online, is titled AI Research Experiences. And this was, I would say, four years in the making. In the back of my mind thinking there needs to be a course for this.

Do they have to publish a paper to get cited 7,000 times to be able to, to get an A? Is that the, is that the bar? No, absolutely not. My interest for a long time has been can we take first time researchers who are talented, maybe know machine learning, certainly know some computer science, and [00:57:00] enable them to do AI research quickly.

And I think the kinds of skills. We learn in traditional classes and the kinds of skills that we get out of a research experience, there's a gap somewhere and people are expected to fill up that gap on the fly. And remember, in the first year of my PhD, I spent time in five different labs. In undergrad, I was in two different labs,

and I understood how that learning process could be very unstructured and therefore detrimental to actually being able to make progress. And then I thought, what are skills that one can learn the form of a class that allows them to say, I can do software engineering. Well, I understand how to work in a group.

I understand how to navigate the reading of research papers, and then I can navigate the writing of a research paper as well. By the time people are finishing their PhD, They're [00:58:00] masters of all of this, but I think the rest of it can be taught in a course format in three months. And that is my hope with CS 1 97.

So it's called AI research experiences, basically covers. If you come out of a machine learning class, what is everything you need to know with an asterisk there on my version of everything you need to know to be able to contribute to research very effectively. And is it primarily undergraduates? It's primarily towards undergraduates, yes.

So it's a different focus than the Coursera course, right? Yes. Which is really to get very quickly up to speed with being able to, to pick up those skills for writing research papers and contributing to research projects. Cool. Alright. Pranav, is it okay if we do a quick lightning round? Let's do it.

I forgot my ear horn though. So horn. Horn. Yeah. So brevity is the soul of wit. So you get two sentences to respond to each of these questions. That will be in rapid fire. Um, so. I always hear [00:59:00] from clinicians, how am I gonna trust AI if it won't explain itself to me? So my question for you is, should AI be explainable?

It depends on the use case. You get one more sentence, . Sometimes you need to explain at a global level how a model works and sometimes you need to explain an individual decision. Expertly done. Pranav, What's your favorite novel? I like the Gene, although it is more a non-fiction book than a novel. We'll, we'll count it.

Isn't it called an intimate history? Isn't that so? Maybe that's true. Maybe, maybe that counts. Given the spotlight that you were put on, during your PhD around human level physician performance for chest xray interpretation, is Twitter a meaningful medium for scientific discourse? I think it is in terms of both discovery

of what's interesting and in [01:00:00] terms of hearing different perspectives on a problem. Okay. Do clinicians need to understand machine learning to contribute to machine learning projects? I think in the same way that users of apps don't need to understand how an app works as much as how to interface with it, I think I would like to put the burden on the developers of the apps to get right.

The user interaction. The idea being the user is never wrong. Will machine learning reduce costs in healthcare? Yes. Well done. Are preprints a net scientific good? Yes. The final question of the lightning ground is if you could have dinner with one person alive or dead, who would it be? I think I'd have to go with someone who is, who is dead.

I'll say [01:01:00] probably Richard Feinman. Okay. Just because I have read a couple of his books and I thought it would be really fun to be able to have a chat with him. All right, cool. Thanks. That's the end of the lightning round. So I tallied up your score. You've made it to the showcase showdown. You've made it to the, the final segment.

Uh, you passed the lightning round. Um, okay. So now we're the last segment, um, of questions that we want to ask you about are big picture and sort of concluding questions. So I guess the, given your perch, your successful history of collaborations with clinicians, and given everything that is happening in medical AI now, how should your typical clinician think about the impact of AI on their profession?

Well, I think it would be useful if there were a way to basically list out the problems. That, uh, you face as a clinician on a day-to-day basis and have your [01:02:00] wishlist for what you consider to be somewhere you would need help. And, this should be a wishlist that is of the following form, which is if I could have a help with any given decision, what is the kind of decision I would want?

What is the format of the decision that I would want in terms of the output? And what is the data that I would expect this tool to have? And I think if as a clinician you have the ability to set out this problem definition and start collecting the data or start to figure out how to get that data, I think it is very likely that you will be able to build a partnership with someone who is able to build a solution for that and start getting a headway in terms of integrating that

in your own practice and hopefully be adopting in the future, a commercial solution [01:03:00] that takes care of that in a, scalable, safe manner. So I would say in terms of, steps to take now, that would be number one. Okay. Just to push on that a little bit more. So going back to the CheXNet tweet from your advisor, do you think over the long term physicians should be thinking seriously about replacement by AI?

I think there are two aspects to that question of will AI replace doctors and let me talk about radiology. I think one of the aspects of that question is capability. If we think about the capability of algorithms to take in an image and output a report, will that capability be achieved? The answer is absolutely yes, and it will happen in the next few years.

And then you think about layers of complexity on top. You say, okay, but I take into account the indication that will happen. I take into account the [01:04:00] prior that will happen. Now the issue is that in a lot of databases, that linkage isn't well characterized or, or clean. And so it takes time as a result of the data not being in the right format.

But in any case in which the task is well defined, and all of the information that is accessible and available to a clinician is made available and accessible to a machine, that decision is going to be made and it's gonna be made really well. . Now, that is a different question from the replacement question because there's also a usefulness component to this, and that is, if you're thinking about a clinician's workflow and about the kinds of things that a clinician is able to do, whether it relates to conversations with patients, whether it relates to incorporating different sources of data, some of which might not be captured anywhere, then that is [01:05:00] something that the role of the clinician would direct towards more.

And so I think the replacement question is dissociating into two questions, which is what is the interaction of the physicians with the tool and then what is the capability of the tools and the replacement question can be thought of as the meeting of these two sub-questions. Got it. Thanks. So we've talked about trustworthiness, robustness of AI. Obviously these are big themes and very important concepts, so I'm curious if you could give us your thoughts on whether at net will machine learning, artificial intelligence, exacerbate healthcare systems' disparities? I think not with the right kinds of education to the community and the right kinds of support for the research that tries to figure out

[01:06:00] how safe and effective and equitable these systems are. I think if we continue to support that research, one of the reasons it's important to support that research is that they may not be incentive to ask those very questions in industry where the questions of intelligence and how we can push that might be more aligned with a profit making incentive while the ability to make systems fair and equitable might not be very aligned there. And so I think the burden will fall onto academia to be able to make those claims. And I think there will have to be an interplay of being able to say, I can try to verify the accuracy of these systems by having some kind of transparency into at least their outputs, if not their internal decision making process to be able to validate whether or not

these things work in practice, and I think as [01:07:00] long as that separation can be kept of not having industry influence whether or not such studies can take place, then I think we will move towards a world in which we're able to have these systems be safe and equitable for all. I think also I want to add to this, and this is one of my hopes, is that there is this natural default that the way we're gonna deploy these systems is going to be companies that are gonna be selling their products

but I envision a world in which it's going to be free services that are tools that are created by people that are going to be interchangeable, one for another, or all working in some sort of a collaboration or a, let's see what the majority of these say. In a way that's very open source that you and I could go and verify what the code is actually doing.

We could [01:08:00] modify it. And this sort of open development format in a way that's happened to operating systems and has happened to other code, as the way we actually go about deploying these systems in this system. How would a regulatory agency. Think about that future. I think it's going to be challenging to think about that, that future from a regulatory standpoint, but in the way.

The FDA regulation is saying we're gonna establish a regulation on the process used to update models and develop that. I think if there's a system which says, for something that is developed by people all around, these are the steps that are going to happen before we're using this on real patients, and those are agreed upon by the regulatory bodies and by all of the different stakeholders involved.

I think that's, to me, seems like a good starting model for how we bring about these models into practice. This, I think might be my favorite [01:09:00] question, and feel free to go as big or small as you want. What is your most controversial opinion? This is an easy one. Okay. I think my controversial opinion is then try harder.

My most controversial opinion is that I think de-identified medical data. should be a easily accessible resource for researchers, and this is certainly not the way things are working right now. As you know, healthcare data is very siloed and privacy is one of the explanations for why this data is not widely available.

But I don't think that's a very good one because I think there is a lot of value to be had in making that data available to researchers and a lot of ways of acquiring consent from patients. If that seems to be the bottleneck of being [01:10:00] able to do that and make that data available, in fact, the data should be available even maybe by default, and the consent should happen by default rather than explicitly doing it per project.

I think that would be a very important model for us to go forward, and I think in some ways, We're already 10 years or 15 years behind of where we would be compared to a world in which we had easily accessible and available data. And this means data, not in just a small sliver of a vertical, like a single chest x-ray associated with some labels is to be able to get a overall picture of of a patient.

I think there are going to be technical approaches to get around the privacy challenge of being able to say, I have sensitive medical data that I cannot fully de-identify . Yet. I'm still able to build useful models off of this. I [01:11:00] think we will work towards this in the future, but I worry that even if we have the technical approaches, we're gonna have just.

challenges in setting up the infrastructure to be able to say we have all of these different data sources that have never talked to each other that are very hard to even go and do individual reach outs to coming together to solve a problem. And so I think if we take the approach of, we all agree that medical data is a resource that we're not going to try to sell, cuz I mean, that is a model that a lot of hospitals take, is this is a valuable asset, we can sell it.

And, you know, in the limit of infinite resources, maybe the right thing to do is to buy all the data from the different hospitals and make it freely available for all. And I think this norm needs to change and I hope that it will change in the, in the next few years. Great. Pranav, [01:12:00] one final question and we'll end on a optimistic note.

What are you most excited about, uh, in the next five years for medical AI? I'm very excited about pushing the capabilities of these systems. I think there was a wow factor in 2017, 2016 when we saw the first examples of these models that were able to perform as well as experts on these really well-defined tasks.

I think the next five years is thinking about how we push that to be able to have more wow factors to say all of these things that we thought were traditionally in the domains of hyper specialized systems or systems that required a lot of resources or a level of intelligence that we didn't think possible or a level of coverage we didn't think possible suddenly become possible.

And one of those directions, as I mentioned, is the generation of reports. But then on top of that, you can think about layering on explanations. [01:13:00] You can have conversations looking at an image between a clinician and a model. And I think that future is going to be possible and is going to happen in the next five years.

Awesome. Pranav, what a pleasure. Thank you so much for joining us. Thanks for having me. Yeah, thanks.

Thanks again for joining us on another episode of NEJM AI Grand Rounds. Be sure to follow us on Apple, Spotify, Google, or wherever you find your podcasts.