AI-GR Pod 39 Seth Hain 02.11.26

[00:00:00] There are many things like new styles of devices, direct relationships between the AI technologies and the scope, as an example. Or new styles of therapeutics and diagnostic approaches that as patients we think are very, very promising. And as a company, we think it's important to be able to work closely with.

And so, we see folks as an example through something we call Aura placing orders in Epic for specialty diagnostics that come from a different company or maybe from a different type of device or a personal device that the patient will be prescribed. And that information flowing back into the physician medical record to be able to drive insights and that type of structure that supports all of these different types of AI.

Where it connects best to be it the test, be it the device, et cetera. And then how do you put it back into workflow? I think there's a role for us and there's a role for many other people in that type of [00:01:00] future.

Hi, and welcome to another episode of *NEJM AI Grand Rounds*. I'm Raj Manrai. I'm here with my co-host, Andy Beam, and today we're thrilled to bring you our conversation with Seth Hain. Seth is the SVP of R&D at Epic, the company that you're likely very familiar with. That good chance makes your organization's electronic health record if you're at a hospital.

I mean, we had a really wide-ranging conversation with Seth. We got to learn about his background in mathematics. How he started working at Epic and his work over the past few years on building AI models using massive amounts of data at Epic. Andy, I really liked some of the shared connections that you had with Seth around interest in math and video games, and then also I think 4-H, which was not on my bingo card for what would come up during this conversation.

It was a lot of fun.

Yeah, agreed, Raj. This conversation was unexpected for me because I thought that we were gonna talk about Epic [00:02:00] as the sort of monolithic EHR that everyone knows and has interacted with. But what I really enjoyed was understanding how they think about research and development. So, Seth is the head of R&D at Epic.

They have a lot of really, really great thoughts about AI, its impact on health care, thoughtful ways for training and evaluating health care models at scale, pitfalls for how that can be deployed. And so again, I found the getting to understand how they think about the more researchy side of things to be super enlightening.

And I found Seth to be really thoughtful and we had a really, really engaging conversation.

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

And with that, we bring you our conversation with Seth Hain on *NEJM AI Grand Rounds*. Alright, Seth, well thank you for joining us on *AI Grand Rounds*. We're excited to have you on the show today.

[00:03:00] Thank you for the invitation. Excited to join.

Seth, it's great to have you here. So, this is a question that we always get started with.

Could you tell us about the training procedure for your own neural network? How did you get interested in artificial intelligence? And what data and experiences led you to where you are today?

So many, so many areas to dive into there. Personally, what led me here, I, my background was starting in mathematics. And so, that's what brought me to Madison, Wisconsin originally was graduate school. And for many years here at Epic, I focused on the backend architectures that supports scaling out the many different applications from MyChart to EpicCare in the exam room to the inpatient capabilities.

And I think a key aspect of that has always been that at Epic, our focus has been on how do you capture a patient's story? And then use that in many different [00:04:00] contexts, and over time that led to many of the insights that ultimately resulted in what we are building on top of from a foundational model perspective on top of Cosmos and the training that we're doing in that context.

And so, obviously many, many steps along that path. But I think an important part of that is pulling together both people with a deep understanding of medical records and the journeys that patients are on, and an appreciation of the data in that context combined with building out a team deeply familiar with the pre-training process and beyond for creating foundational models.

So, that's an abbreviated answer to your question. I'm happy to dive into the details.

Yeah, so you're skipping ahead a couple chapters there. I'd love to hear a little bit more about what made you interested in mathematics. How did a mathematics major end up in an EHR company? Were you interested in medicine?

Could you take us to some of those like [00:05:00] formative data points and help us understand how you got to where you are?

Yeah, happy to do it. My background in mathematics came from a very young age. And I think as many of us probably got to experience, I had a phenomenal teacher in high school that really laid the foundational interests around mathematics.

Combined with kind of as a fun aside, I grew up outside of Lincoln, Nebraska. And I was the odd kid whose dad created a computer science program for 4-H. If you're familiar with 4-H, it's a program where kids tend to raise cattle or pigs or bake cookies and take them to the county fair. I was the kid that brought his printout of a computer program to the county fair.

So, I had this combination of at-home programming with a very strong set of teachers around mathematics. That led me to sort [00:06:00] of this combination. I found myself in Madison for graduate school and there was a company that was starting to grow. They had just signed Kaiser Permanente and I thought this could be an interesting opportunity.

And in particular, and this, this remains to the day, Epic is willing to hire folks straight outta school that have a combination of interest and curiosity with an aptitude for quickly learning new ideas. And so, I was lucky that folks took a chance on me in that context and, uh, somehow that has led to now 20 years here.

I took a 4-H magic camp one summer, so I can also attest to the impact that 4-H can have on kids early on. Were there other things, so I know that Epic has this large presence in Madison, Wisconsin and is really great at pulling from like deep quantitative talent in that area.

Were there other things that you were thinking about at the end of your master's program? What were the set of options [00:07:00] that you were thinking about, and then sort of how did you decide on Epic as your destination? Obviously, it's been a good home there. You've been there, as you said, for two decades now, so obviously you made the right choice.

Yeah. Yeah. At that point in time, there were three different routes I was considering. One was certainly staying in academia, a route that has a lot of opportunity. But I tend to want to and understand a wide variety of spaces. And working in industry, and in particular here at Epic, I have the opportunity to move between areas and continue to grow my breadth of experiences across multiple different domains which was appealing here.

The other thing I was very interested in was working in an industry where you could help people. And you could do so in a collaborative, productive way across industries and including with academia. And Epic met that criteria as well. And so, it was an opportunity to work in an industry, solve really hard [00:08:00] problems, and do so in a way that my family members could benefit from, as well as community around, now, the whole world.

Obviously, that scale was something that happened over that time period, though.

Nice. So, we're gonna talk about what you're working on now in just a bit. You showed up at Epic, you're now, I believe, Senior Vice President of Research and Development. How did you go from Madison, Wisconsin grad to SVP of R&D at the world's premier EHR company?

It was a journey that involved both learning from a number of folks here and taking opportunities as they came. I started by focusing on very low-end scalability, optimizing all the way down to the chip set, all the foundation behind Epic, something we called Chronicles. And over time, that led to working on how do we virtualize the architecture?

How do we build out a cloud-based platform for our [00:09:00] applications to be deployed in essentially a platform as a service that runs on the hyperscalers, as an example. And working through those types of steps, building out a series of products that ultimately included building out the artificial intelligence platform or something we call our cognitive computing platform.

Interestingly, I was offered the opportunity and took it to gain familiarity with our clinical applications. It's always good to bounce around and have direct experience with end users using the software, particularly both the clinical folks but then, also, the back-office folks. And so, in January of 2020, I started working across the suite of applications that are essentially everything you walk in for from a, office visit, to the urgent care, and ED even to, and this, this was fun, going on immersions to dental offices for our Wisdom product.

Obviously, January of 2020 was an interesting time to start [00:10:00] working with that group. It was an established set of products that suddenly we needed to think about, for example, which clicks in the flow for administering a vaccine are deeply, deeply necessary. And how would you do that if you had a line of cars that needed to happen?

Where historically the vaccine administration activity was very focused on children, uh, during well child checks. And so, you end up going through those types of pieces as well as getting deep into analyzing things like lab results and those pieces around the COVID tests, as another example. And then in late 2022, we started down the path of rapidly rolling out generative AI first in the context of outpatient care.

And that led to something we can kind of talk about the trajectory of in more detail, but an intense focus on how to apply generative AI to first help [00:11:00] clinicians directly helping patients and, also, doing a variety of different research activities to advance medicine. Um, which has been a consistent focus of mine over the last couple years.

Cool. Thanks for that, Seth. So, I think that’s a great segue to hop into some of the work that you’ve been doing at Epic recently. In particular, I’d like to focus the discussion on a paper called Generative Medical Event Models Improve with Scale. We talk about scale a lot on this podcast. I’m excited to dig into how you guys have been thinking about it.

First, maybe could you set up this paper? What were the goals? What were the key findings? And then maybe also what were some of the key challenges?

I think it may be useful to take a step before even the paper, just so the audience has full context of Cosmos, which was a foundational piece that opened the door to make this paper possible and then dive into the findings, if that’s okay. Because I think that the context is useful here.

Cosmos is an effort across the Epic [00:12:00] community to bring real world evidence into the point of care in the context of that visit between a patient and a physician, as an example. This is an effort that has been underway for almost a decade, at least at this point, and started with a focus on building out a de-identified data set for the purposes of doing real world analysis.

Eventually, in real time, for example, in workflow. This effort is governed by a variety of health systems and individuals from the health systems that are participating are elected to a board that creates the governing council for Cosmos. And Cosmos these days has over 300 million unique patients in it.

When I say unique patients, what I'm referencing is that if an individual goes to multiple different health systems that are part of the Cosmos community, their medical record is stitched together across [00:13:00] those. So, you have that full patient journey, and that's what the de-identified record represents in the Cosmos system.

And so, the first phase after starting to compile that was that health systems participating in it, in Cosmos, had access to the dataset for purposes of research. And there's been over 130 studies published in over 80 journals built off of that data set. In parallel to that, we started to build it into software as well.

That included, for example, connecting physicians that might be seeing patients with a similar constellation of symptoms so that they could discuss that kind of unique situation that was presented to them through the patient with a colleague, and identifying those folks, which you can do through this.

It also meant generating growth charts for kiddos with kind of a unique, say disease. Not truly unique, a rare disease, if [00:14:00] you will, outside of the traditional growth charts. So, them and their parents could see how they compared to other kids in a similar circumstance. And it was those types of applications that led us to wondering if there was a way to make it faster and more efficient to build applications using the Cosmos dataset and do so in a way where it would apply to essentially any diagnosis or a patient at various stages of their medical journey.

And make it far more applicable to the broader health care community. Can I hop in and ask a question about Cosmos? My experience in health care and with hospitals is that they are reluctant to share data with other hospitals and other institutions.

But it sounds like if you're using Epic on the backend, then you're contributing to pooling what is by far the largest health care data set in America. You know, one of the largest in the world. So, what's the operational [00:15:00] side of this? How do institutions contribute data? Can you access data from other institutions? Maybe you could just talk a little bit about that.

No, it's an important aspect of this, and I would say that foundational to the entire Cosmos effort is trust. And one thing I want to be clear about, while there are organizations that use the Epic software that do not contribute in participating Cosmos, it is rapid and straightforward for an organization when they're ready and they decide that it makes sense to join.

And that includes both a mapping effort, as you might expect, and then the transport into Cosmos is pretty straightforward from a software perspective, after we've engineered that. The key there is an understanding that that data set and the rules of the road that are governed by that elected committee of health system professionals is designed to support in workflow real world evidence.

There is no secondary data use of [00:16:00] Cosmos. It's designed for those health systems participating in it to both do research and then use it at the point of care. And that's what establishes the trust.

Cool. Thanks. So, I think we were to the point where we have the database now, and now we can talk about the big model that you guys, right.

That's right, that's right. We, I felt like we needed to set the foundation before we got. That's very, very good context. Yeah. Thanks.

That was great context. And I think the key here, in regard to the model we train was the recognition, as I noted earlier, of the value of a patient's story. And in this context, usually folks think about stories or training models around natural language.

A story they might read, a novel they might read, a Reddit post they might see. And how in pre-training a model tries to predict the next word in that story or that Reddit post and then compares it to the actual ground truth of what was in the dataset that it's being trained on. Right? [00:17:00] That's the general path of a pre-trained model.

We identified that you could tell a patient's story, and others have started down this route as well, which are referenced in the paper, through a series of structured medical events. Meaning the diagnoses, the orders that are placed, the observations and medications that are part of the medical record, and importantly, the time that passes between those.

And rather than predicting natural language words, the question was, could you predict the next event in that patient's story? And I think this was really, I just wanna emphasize this point for the audience. When we start training one of these structured event models, there is no foundational knowledge of what is hypertension.

Or what is Lipitor? They have no idea what that is. The tokens are directly [00:18:00] mapped to these structured medical events in the vocabulary and then trained on that sequence of events, and that is what led to us training what is now the largest structured medical event model in the world. That training cycle ended up ultimately with a billion-parameter model that used over 115 billion discrete medical events in its pre-training process.

The paper that you referenced has kind of a, what I tend to think of as three major sections, and we can dive into those. The first one is around the series of ways to evaluate a model. We ended up building out 78 different evaluations for understanding how a structured medical model could impact everything from diagnosis, say in the near [00:19:00] term as you're going in for, um, an office visit to prognosis over, say, a year or three years of a chronic disease, to things like operational metrics around length of stay, readmissions, your likelihood of going into the ED. And so, we built out a large eval suite.

Obviously, another step of that is then the actual training of the model. How it's structured, how you do the pre-training, et cetera, which is explained in the paper. And that, importantly, as loss drops, when you're doing the pre-training, it correlates to the impact on those evals that you want from a medical perspective.

That's what you would hope for. Um, but it's an important thing to evaluate. And then the third piece of the paper, and maybe we want to dive into it, I think it's ultimately going to prove to quite possibly be the most interesting piece, [00:20:00] is that we studied the scaling laws in the same way that people have for natural language models on these structured medical event models. And found that in a very similar and predictable manner, as you scale up the amount of data and the parameter spaces of these models, you can improve the quality of the models.

Meaning that if we or others continue to train larger and larger models, they'll get better and better at those types of evaluations we talked about.

Awesome. A couple things I'd like to follow-up on there. So, maybe in the context of the evals, could you talk about how you would imagine this model being used in clinical practice?

I have a couple things in mind, but I'd love to hear at least what you guys were thinking as you trained this model.

Oh, there's, there is so many and I think one of the things I'm very, very excited about [00:21:00] is opening up these models to the researcher community on Cosmos, which is planned for the first quarter of next year to see how they do this also. But lemme give some specifics.

I think the first place we'll see this model being applied is in many of the contexts where folks build traditionally predictive models for a single given target. An example of that might be predicting and identifying individuals at high risk of type two diabetes complications over the next three years where this model auto regressively predicts what's to come in that patient's journey, including those complications resulting from diabetes and then use that in workflow to help prioritize care manager outreach, as an example.

This is something people build predictive models for today, but in the context of this model, you can predict many different chronic [00:22:00] diseases as well as the possibility of co-occurrence between different types of outcomes in that context.

So, I think that'll be one example, one that I'm very excited about that we looked at, um, quite intensely with those evaluations was what is going to immediately happen as the patient walks in for the office visit? What are the likely diagnoses in this context? What are the likely medications that will be ordered, et cetera.

Obviously, that has a number of different ways it can surface into workflow. One of those is through something we're calling diagnosis checker, as an example, which is essentially both a differential diagnosis essentially that can be generated with a tool like this as well as likely subsequent labs and tests that can be done to help determine if that diagnosis is accurate or follow-ups based on it.

And so, in that context, we see the model helping generate and [00:23:00] suggest, as it makes sense, always have with the human physician making the decision at the end of the day, but considerations based on that large data set and doing so in a way that is efficient within the software. Which is always part of this is how, how do you help increase efficiency?

The third example of this, and the third kind of category of specific cases tends to really be around hospital operations. Given the volume of challenges facing the health care system these days, and in particular the aging population, which needs increasing amount of care. And candidly, the limitations in regard to access that many health systems, hospitals and clinics are facing right now.

Being able to accurately predict things like length of stay, being able to understand the likelihood that somebody will need a more intensive level of care within the hospital, et cetera, is a meaningful set of data points to be able to help plan out everything from [00:24:00] staffing to occupancy rates, et cetera.

And I appreciate that in many of these contexts, the things that I'm saying are things people have built individual predictive models for one at a time. Those are intense efforts from a data scientist perspective that includes building out those data sets, et cetera, et cetera, which I'm sure the audience is quite familiar with, but are ultimately around a single target that they end up predicting.

Our purpose with building out this generative model was that one model could do all of those predictions and open the door to more complex investigations in the future.

Can I ask a question just to make sure I understand correct? 'Cause you have a model trained on 300 million patients and presumably hundreds of different hospitals, academic medical centers, local clinics, anywhere that they're using Epic and they have opted in.

Is there a way [00:25:00] that the person can say, what's going to happen to this person at my hospital? 'Cause it's very hard for me to think about the trajectory averaged across hundreds or thousands of different institutions because all of those are gonna have different policies, different ways of treating.

There's gonna be a lot of, like, heterogeneity, both an outcome, but how they treat patients. And so, sort of thinking about the average trajectory of a patient in a vacuum is, is a hard trajectory for me to think about.

No, it's good question. And a couple of different pieces here that I think are, are worth noting and, and one thing I'll just quick note.

Cosmos, the large data set, has 300 million unique patients in it. Um, the largest model to date, we've trained on Curiosity, has 118 million patients in it. Purpose of the scaling laws is it shows we're gonna keep scaling up that compute in the volume. So, just a, a brief aside on, on that specific piece. As you note in this context, the current version of Curiosity, or what was formerly known as Comet, in [00:26:00] some context, is essentially a way to generate real world evidence across the large population in regard to what are the likely events to follow in this patient's trajectory.

So, when applied in the context of an individual situation, you start with their medical record story and then auto regressively predict the next set of tokens as they come. Now in practice, what this actually looks like, and we can get into this in more detail, is you actually generate that multiple times. The result of this is that you have 20, 30, 40 potential timelines that this patient can go down, which gives you a sense of the odds of different types of outcomes that can occur in that context.

Now, as you noted, this current version of the pre-trained model is across to the full population. It's similar in some [00:27:00] sense, to doing an analytics query, if you will, against the Cosmos dataset to understand outcomes of similar patients in that context. As an example, we then use software capabilities to localize that to the individual's individual environment in that context and put into place their practices, et cetera, which can inform and overlay that type of additional information.

We are interested in the future from a research perspective. What things, like, what post-training techniques and fine-tuning techniques on this model might be applicable in the context of an individual organization to also account for some of those, but there's different techniques to address that localization, if you will, of the general model that will be both deployed within the software, but then also from a training and research perspective afterwards.

Super helpful. I think that like localization I think makes [00:28:00] a ton of sense. One more question and then I'll throw it over to Raj. My sense is that doctors often don't want predictions, that they want decisions. They want, okay. If this patient is headed towards this outcome, what should I do to steer them away from a bad outcome? And how do you think about that use case? Because my, like, epi- epidemiology sensors start lighting up and like confounding and like, how do you actually recommend an intervention for a patient given that those decisions often require very careful analysis, very careful setup to control for things like confounding.

Uh, this is a great and important question and, in part, it gets to how we intend to roll this out. And as I noted earlier, I talked about Cosmos in general and how the first phase of this was getting researchers access to such a large data set for writing papers, exploring how they'd use it in that context. And in [00:29:00] a similar way, with these Curiosity family of models, our first phase is going to be giving researchers access to that, and that includes a visualization tool. So that as a researcher, you could imagine this being for a physician in the future in the exam room. You actually see these 20 trajectories that a patient might go down in commonalities across them so that you can investigate it.

I talked to physicians and we require all developers to go on immersion to see their software in action. They often think of and appreciate that individual events in a patient's kind of future trajectory might send them down a different route than anticipated. And they can see that through these multiple timelines that end up getting projected.

Now, that doesn't fully answer the question, Andy, that you asked. I think that what we are hoping, and what we're trying to encourage [00:30:00] is that by one, publishing this paper and getting it out to the world, and two, creating a set of tools for folks to explore it, they become more familiar with this type of evidence at the point of care, which ultimately leads, along with working with HCI colleagues and others on the best ways to embed it into workflow for everyone to be able to consume.

Because you shouldn't have to be a researcher to have an appreciation for this in workflow, but I think there is a need for the science to progress around this and for the education to happen, and we're trying to create an environment to support that.

It's super helpful, Seth. So, maybe I can actually just before we zoom out from Curiosity and Cosmos, I can just ask one follow-on to one of Andy's questions. This is about localizing the model and understanding how well it does at a specific institution or how well it performs across institutions.

I know it's a, like, ton of work just to, to train the model [00:31:00] right? To do what you guys did in releasing the pre-print and the various incarnations from the small to the large model and so on. But I'm curious if you, in the context of that paper or if you have planned future work, if you've looked at just empirically the performance of the model across different types of institutions.

Across different sort of clinical settings. And if you have a sense of where and when the model tends to do well, and maybe that can be somewhat defined by institution type as one of the variables, but maybe it's maybe you have another sort of natural variable you would go to, like the length of the patient's history or the duration or the frequency in which we can sample their data over the last six months or year.

How do you think about where the model is likely to do well and where it's likely to falter or have a kind of wider and less useful prediction for a given patient?

No, good question. So, on the specifics around localization.

This is a question of [00:32:00] deep interest to us, but not one that we have fully researched yet. So, it's something under trajectory and we have capabilities within the Cosmos dataset to be able to compare and contrast this across different types of institutions. I think it also gives us the opportunity to help make sure that patients, regardless of where they're getting seen. In some ways, growing up in rural Nebraska, I'm excited about the opportunity to use these types of models to help folks in other situations get better and better care.

And I think the type of question you're asking also leads to quality improvement programs that can elicit less well. How does it work on different organizations and more importantly, what are the differences that make this work really well in certain organizations versus others, and how can we share that knowledge across?

One of the things we're very interested in right now, and folks that all have read the paper will have noticed this, is [00:33:00] both expanding the model to additional populations. An example being kiddos.

We, we have the opportunity. So right now, right now it's just adults? It's just adults. Just adults. It's over the, the population we trained, the first model on was over 18 years old and as we are ramping up for additional training runs, we will be including pediatrics in there as well, which, I think is going to be important for a variety of reasons. We work closely with our pediatric institutions across the country and look forward to working with them on research around this. So, that's one piece.

The other thing that we're working on is expanding the context window. For audiences not familiar with training transformer-based models, what that basically means is we want to tell more of the patient's story in regard to the prediction, and right now we have a limited window in regard to the number of structured medical events that can be added.

Obviously, we tell them up to the recent [00:34:00] pieces, so continuing to expand that context window and then also adding in additional types of demographics data as an example in expanding the vocabulary. So, I think those will be some of the key pieces you'll see us focused on. In addition to obviously continuing to apply those scaling laws.

And maybe as one last sort of follow-on Curiosity. I think in the paper you make a lot of comparisons as one of the baselines, both within the different sizes of the models, right? Towards this big theme around scaling. But I think very interestingly and importantly, you also compare it to supervised versions of the same setup, right?

Where you try to have the same predictors and then ask how well, like a supervised model would do on the same task to see if this model's learning something, that there's more sort of traditional approaches to the same prediction challenge would not. And I think there's a few different comparisons.

There's different types of supervised models that you're using. I'm [00:35:00] curious, one of the models that we've heard a lot about, even before this new and interesting work, was your efforts for predicting sepsis for predicting other types of events. Did you do a direct bake-off between the Epic sepsis model and this one?

Um, I don't know if I caught that in the paper. I'm curious if there's plans to do that. It's very interesting 'cause there's like kind of competing philosophies here, right? For how to actually best predict the clinical outcome and what's also even to the point of localization. What's likely to be robust enough to survive across different institutions, different ways of coding, different environments as a clinical feature that's stably represented versus something that's institutionally patterned or shortcutted at a given site.

So, did you do a bake-off with the Epic sepsis model or would you do that if you haven't done it already?

Great question and I think it gets to a key aspect that we have yet to do with Curiosity. The short answer is we have not done it yet. The longer answer is that a big part of using [00:36:00] models, be those traditional predictive models or these generative AI models, is how you configure them into workflow and how they relate to the specifics of that organization on say, that floor of that hospital. And so, in regard to things like a deterioration index model or a sepsis model used in a production context, a large part of that involves who those predictions go to. and at what points from an operational perspective.

And then measuring outcomes as a result of that configuration in that context. And given that we haven't deployed Curiosity in any circumstance yet, or in that context, we don't have a comparison point at that level, and that's where a lot of those different questions end up coming up.

I wanna be clear though. It'll be important as folks begin to explore Curiosity and deploying it in [00:37:00] different contexts, that they do have that baseline of impact on the existing workflows. And as we do with all models that can be rolled out across the Epic suite of software, there's an opportunity to run the model silently, behind the scenes on the local population in those workflows to understand those comparison points before somebody puts them into practice. And so, we would see that route being what we would head down in this context to answer the important question you asked Raj, but we don't have that answer today.

Got it. So, maybe we could just zoom out a little bit from Curiosity and from your Cosmos work. I imagine that in your role you're getting pitched both sort of internally and externally all the time about. The greatest new generative AI product to integrate into Epic's electronic health record. And I know that there already are several efforts, right?

And rollouts of generative AI products that are integrated with that bank. And so, I'm gonna ask you, what I'll just say at the outset is an impossibly hard [00:38:00] question. You can answer whatever you could answer of it because I don't think anyone has the answer to this question, including regulators who are trying to figure this out, but from your perch, from your perspective at Epic, how do you evaluate the trustworthiness of generative AI products that folks would like to integrate into Epic to serve your users?

I think this question of kind of trustworthiness and evaluation, and in particular the history of where we came from, from a machine learning perspective, is a conversation that needs to continue to evolve. And there is a baseline set that needs to be understood around any model or any tool in regard to its accuracy in the context of how it will be used.

The interesting thing that we're starting to encounter when applying generative AI into workflows is that it can do [00:39:00] things at scale that we never thought it was worth putting time or effort into to do, due to limitations of staff, et cetera. I'll give you a very simple one from an Epic example. So, we have a capability in the software that generates summaries of medical records.

It generates different types of summaries in different contexts. Obviously, you want a different summary if you have a patient coming in for a physical after five years, than a nurse wants at the end of a shift, as an example. And it can generate those different types of summaries. Now, no one was writing those summaries before, so you didn't have a baseline of how to compare against that.

And so, in that regard, we've ended up working with a variety of researchers, including some folks here at the University of Wisconsin, to build out a series of evaluation techniques, including ultimately LLMs as a judge. All of which is part of the, what we [00:40:00] call the AI Trust and Assurance Suite, which is an open-source tool for this, and it starts to answer some of these questions you couldn't answer otherwise.

And I think that folks need to take that extra step back to think through what is the ground truth that you're comparing the use of these tools to when thinking through whether or not and how you need to deploy them and how to measure their accuracy and appropriateness in those contexts. It's quite a bit different than, say, historically I reached out to patients based on ordering their A1C lab values, and now I have a predictive model for their numbers of diabetes complications.

How does this compare? You just don't have the same sort of baseline.

Awesome. I think, great discussion, and I think this is a great time to take a quick break and go to the lightning ground. How's that sound, Seth?

Sounds good. Looking forward to it. [00:41:00]

Awesome. So, uh, if you haven't listened to the podcast before, these are, uh, short and sweet questions that usually get short and sweet answers. You get to decide which one to take seriously and which ones to take unseriously, but we'll just, we'll run through 'em. Alright? Let's go.

Alright, so the first one is, one has become one of my favorite questions and it is, what was your first job?

Delivering papers.

Oh, nice. Paper. Delivery boy, newspaper boy. Nice. Cool. How old were you?

Oh, I think I was in fourth grade.

Oh, um, yeah. Amazing. No, right. You, you've just, you've got all of the, like, Americana, 4-H, paperboy, like going on. I had the bag, I had the bag over my shoulder and roll the papers in the basement with the rubber bands.

Nice.

I also simultaneously played Paperboy on the old Nintendo. I was gonna say, that was my main, remember that? Yeah. Yeah. Uh, running away from the dog, hitting the zombies, doing the dirt bike at the end of each level. Was, was, was good time. So, yeah. [00:42:00]

Amazing. Alright, Seth. Next lightning round question.

What's a technology you think health care is overestimating right now?

Genomics. And, and I think in a narrow sense it's being overestimated, but in a broad sense it's being underestimated and the real possibilities of kind of personalized therapeutics. And where this all might be going, particularly when combined with kind of purpose-built foundational models around biology, I think has tremendous impact.

But I think in some ways it has felt like genomics kind of into the layperson perspective is underperforming.

Yeah. Nice. Um, what is your favorite place to visit or go on a vacation?

Oh, my family and I have been going to Portland, Maine the first week after the kids get outta school. And we [00:43:00] just love it.

And, and in part it is the food, and in part it is the walks on the beach with the kids to hear about kind of what they're imagining for that summer. So, it, it, Portland's a great spot for folks, but I will also, I also gotta admit, I just got great memories of spending time with family there. So, that's immediately what comes to mind.

I will second the Portland, Maine endorsement. It's, it's a great place.

Yeah. And Seth, I think that's actually where we actually first met in person. That's right.

That's right. The RAISE meeting, too. The RAISE, the RAISE meeting was excellent. And another example of why I like Portland.

Excellent. Alright.

The next question. Will medical AI be driven more by computer scientists or clinicians?

The answer has to be both. I'm sorry. I, and I think that the reason I say that is computer scientists play a big role. The attention mechanism in transformers and how that helps understand the different parts of a patient's story in regard to predicting [00:44:00] what comes next, right?

Which was, which is in part what the paper is about is a computer science technique. Um, but I, I think, we really, really need to think through the ethics of what's being built, how patients will use it, how physicians will use it, and how that will integrate into society in a meaningful and helpful way.

And that needs to be led through joint conversation between deep researchers on both sides.

Awesome. Um, so Epic. The Epic campus is normally described as, like, this Willy Wonka-esque, um, place full of, like, super creative people and it's like got all this quirk, uh, to it. So, I guess my, uh, next question is, what's the most unusual thing about working at Epic?

What's the most unusual thing? You know, what's strange? You mentioned the campus.

Yeah.

One of the things that is incredibly weird is that having been here for 20 [00:45:00] years, you walk by all of this wild stuff, but you, you sort of, it becomes normal.

Yeah.

Yeah.

That's, that's the really weird part. Yeah. And one of my favorite things is to bring family members here, and—. So they can give you another lens on that?

Yep. I get, I get, that's exactly right. I get forced to see it through their eyes and it's so much fun to see stuff again for the first time. Um.

So, the, the most unusual thing is that it seems normal to you, uh, but it, most people find it highly unusual. Yeah.

For folks that haven't been here, I'll share my two favorite things on campus.

One is a staircase that is a Fibonacci spiral.

That's, and then the other one, we had a deli in town that for 40 years, if it was your birthday and you were a kiddo, you would go there and you'd get a free ice cream cone and you'd get a ride on a carousel. And eventually the deli, Ella's Deli, [00:46:00] closed and everybody in the community wondered what would happen to the carousel.

You can guess where it ended up. Amazing. Amazing. Nice. Very cool. Alright, our last lightning round question, Seth, will AI second opinions be routinely available in the EHR in the next five years?

Yes.

Alright.

I think that will happen. We, we have to figure out how, and we need to work together as a community. But I think that will happen.

And honestly, five years feels pretty long on like today's timescale, right? So, yeah, you made that question easy. I feel like I made, I, maybe I should have made that months. But, uh, anyway. Weeks. Uh, yeah, weeks. Seconds. In the next five seconds. Seth you've survived the lightning round. Passed it with flying colors.

That was, that was great. We actually just have a few last sort of big picture questions and I'll start off with the first one. So, we want you to sort of give us kind of both sides of the coin here, and I think you have a uniquely, uniquely sort of qualified position to, to answer this question. [00:47:00]

The first side of this is what's the strongest case for the EHR being the portal for medical AI? And then the flip question is, what's the strongest case for the EHR not being the portal for medical AI?

Yeah, I think there are multiple pieces to this question, but for the moment I'll sort of take a simplified definition of medical AI. And there's sort of two key aspects for the strongest case.

One is the ability and speed of dispersion, the speed of use. It's an important piece in helping scale out the amount of care as a country that we provide. And I think that that will lead to, and we're seeing it lead to very, very rapid adoption. We have over 440 health systems using generative AI in production at this point.

The other thing on the poll case, if you will, on the health care technology side being the lead is in Epic in particular, is around the scaling laws that we talked about with the [00:48:00] medical event models. The scaling laws indicate to me that there is something relatively foundational in regard to a patient's story that ends up getting told and ends up creating an opportunity to ask questions in new ways and to understand those trajectories in new ways in the similarities across them.

And I don't think we know what that truly means. If the scaling laws didn't hold, I wouldn't be as bullish on that. I think the best case against is actually what's happening in some of the biological spaces. In particular, if we end up in a circumstance where from the principles of basic science one is able-based on labs or images or other capabilities to take essentially a point in time of a patient and create treatments.

That ends up becoming a very powerful part of the way health care will be [00:49:00] provided. And I think that that there is a possibility that both of these things end up happening, but I do think that that ends up causing a lot of focus into the direct therapeutics and personalized therapeutics.

Cool. Maybe just one more follow-up question.

Epic seems to have a lot of things going for it in the age of AI. Cosmos being, like, one of the most unique health care data sets there are. I'm sure that you have a lot of computing power to help train models on that. I think one of the things that has changed for me having been in health care machine learning and AI for 10 years now, is that there are lots of really attractive business models and products.

There are companies that are creating auto transcription services where it can record a patient-doctor interaction and turn that into structured data. There are companies that are creating auto-billing things to help with medical coding. Is there any reason why that's not all just Epic in five years?

It just feels like you guys have all of these [00:50:00] resources where you could offer a suite of services across the entire health care stack that would be able to service all of these markets all in one shop. So, is that how Epic is thinking about what AI looks for you guys going forward, where you're kind of a one-stop shop for all things medical AI?

Or are you, as Raj said, fine to be able to sort of be the base layer that people then build on?

I think the answer ends up being both in many ways. One thing I appreciate about working here at Epic and about our culture, and you'll see this on the walls of the bathroom. If you look at our commandments, which are in every bathroom across the entire campus, is a lot of those are focused on us maintaining our focus on where we can have the most impact.

And the scale and breadth of what can be done with AI, obviously has an impact directly in the applications we've created today, and it will continue to [00:51:00] create more of in regard to helping patients and physicians take care of one another, right? And do so in an efficient way throughout the health system. But there are many things like new styles of devices, direct relationships between the AI technologies and the scope, as an example, or new styles of therapeutics and diagnostic approaches that as patients we think are very, very promising.

And as a company, we think it's important to be able to work closely with, and so we see folks as an example through something we call Aura placing orders in Epic for specialty diagnostics that come from a different company or maybe from a different type of device or a personal device that the patient will be prescribed.

And that information flowing back into the physician medical record to be able to drive insights and that type of structure that supports all of these [00:52:00] different types of AI, where it connects best to be it the test, be it the device, et cetera. And then how do you put it back into workflow? I think there's a role for us and there's a role for many other people in that type of future.

Awesome. Well, Seth, I think that sounds like a great place to end. So, uh, let us thank you for coming on the podcast today. Really enjoyed the conversation, and it was really fun to learn about all the AI stuff happening at Epic.

I really enjoyed the conversation. Hope to catch you up on the next set of papers as they come out as well sometime.

Take care.

Fantastic. Yeah. Thank you, Seth. Thanks, Seth. That was great. Bye.