Transcript Details
This is a transcript of an educational program. Details about the program and additional media formats for the program are accessible by visiting: https://reachmd.com/programs/the-convergence/science-exceptions-safety-bar-clinical-ai/54786/

ReachMD
www.reachmd.com
info@reachmd.com
(866) 423-7849

The Science of Exceptions: Setting the Safety Bar for Clinical AI

Dr. McDonough:
Welcome to *The Convergence* on ReachMD. I'm Dr. Brian McDonough, and today I'm joined by Dr. Suvrankar Datta. Dr. Datta, thanks for being here.

Dr. Datta:
Thank you, Brian, for having me on your podcast.

Dr. McDonough:
Dr. Datta, before we get into the technology, tell me a little about yourself. What brought you from radiology and clinical practice into responsible AI evaluation?

Dr. Datta:
Yeah, Brian, the reason I joined medicine was to be able to impact a lot of lives. That's why I did my medical school, and then I decided that I really was fascinated by healthcare technology, and AI was coming up in a very strong way. And, Geoffrey Hinton, he had mentioned that radiologists are going to be out of jobs because AI is going to do all our jobs, right? And that made me feel, “Okay, if I really want to do my work in AI, I should be probably doing it in a field where I can be the person who builds these AI tools that can help radiology, which is less penetrated in a country like India, to all the masses, right?” That was my motivation for being a radiologist.

And after my residency, I just figured that I am really good at being a scientist, probably much better as a physician-scientist than a radiologist. So after practicing for a year or two, I decided to move full-time to build my own healthcare AI research lab. Radiology research is still something that I do and that accounts for almost 50 percent of the projects which we undertake. But alongside that, we also do large-scale responsible AI implementation as well as evaluation.

Dr. McDonough:
That's interesting. You looked inward and saw your strength and thought, "How can I maximize them?" Which a lot of us in healthcare have to do, even when we choose our various specialties. “What do I like? What do I think I'm good at?” That's exciting you chose this, and we're lucky you did.

What is the problem you're trying to solve in healthcare AI evaluation, big picture?

Dr. Datta:
So Brian, my research group is called The Center for Responsible Autonomous Systems in Healthcare. And what we are essentially trying to understand is that, as autonomous AI agents gradually start entering our systems, and many of these chatbots and frontier AI labs come and claim that their models can diagnose patients better than doctors, there needs to be neutral, independent institutions and research groups which are actually creating these evaluations and benchmarks, which show the community—both the patients and the doctors—the real truth. Is it really how these big companies come and tell us that they are absolutely better than all doctors? And are we in an existential crisis? Or is it that these companies choose only very few specific, curated datasets and show the response of the chatbots and the outcomes in those datasets, right?

So there was a very urgent need when I started in the space to have a very neutral, independent, credible, benchmarking research group, and that is how we essentially started doing large-scale healthcare evaluations. We partnered with institutes all across the world, including Harvard and Stanford, as well as in India as well, to really bring out the real-world cases we see, and also to ensure that we are able to do a very strong benchmarking of not just the frontier models, but all different models available to the patients as well as to

the developers who are implementing these in our healthcare systems.

So this is how we started off, but now, we do much more than evals. Evals still inform the majority of the work which we do—radiology evals, healthcare AI evals, evaluating, ambient AI scribes. The scribes listen to the conversations and populate your EHR, right? Alongside that, we are also looking at evaluations of clinical decision support systems for medicine and surgery.

But alongside that, we are also trying to build the data infrastructure for India. And we are trying to connect these large-scale hospital systems who have the data so that we are able to create really diverse and robust benchmarks from real-world patients, which these frontier models that are still trained on Western data have not really seen, so that we can really tell if those models are able to reason like a clinician. Or is it that they have just seen those kinds of cases in their pre-training corpus, which is why they are able to actually solve those cases?

Dr. McDonough:
A couple of interesting things I take from your thoughts. First of all, I think it's good we tend to just look at model performance. I know in my world, we looked at EMRs—what's it doing? But nobody really looked at the guts of what's behind it. And in this case, with AI, evaluating whether you're really getting correct information.

But in addition to that, you said something else which I find very interesting, and it's not just India, but what you're doing is you're looking at real world. You're looking at use of this because, here in the United States, practicing in Wilmington, Delaware is different than Philadelphia, which is different than New York or California. And medicine's always local, and there's so much of it. When you take the entire world, for us to assume that what's happening in the United States with the way the rest of the world acts, or India is the way the United States is, I think you're looking at it big picture.

Dr. Datta:
Oh, of course. The idea is to really work globally across institutions and build benchmarks to showcase how these models perform on different diverse populations. And what we are also seeing, very interestingly, is that the model which works great in the United States doesn't work that great in India. And the models trained on Indian datasets don't really work that great in the United States, right? And as we do more research on this, we will be very surely finding out more insights about the reason why models behave so differently. Because a radiologist who is trained in India and a radiologist in the US are very unlikely to see a pneumonia on a chest X-ray and interpret it differently, but AI models are actually doing that.

So very interesting things we find out once we start doing evaluations and go deeper, and which really shows that AI models are still not at least at the human level.

Dr. McDonough:
One question I have is, referring let's say to India right now, it has a population at a scale, few health systems can even imagine. What does that scale reveal about limits of AI tools that have been built for smaller, less expanded populations, less diverse populations?

Dr. Datta:
So there are around 1.5 billion people who live in India. We have around 28 states and more than 100 different local, regional, culturally distinct populations. And what we have been seeing is that the AI models, which are often trained even within India, let's say a place which is in North India, like Delhi, does not really work well when you try to deploy that in South India, like Tamil Nadu.

And this just shows that the current architecture, like convolutional neural networks, transformers, which is what most of these foundational models use in the back end, are not really great at generalizing understanding of diseases or interpretation of concepts. They try to remember how some pattern looks, and only if that pattern is exactly same will they be able to tell it is correct.

But at the same time, if you are unable to realize the inherent concept a doctor understands by looking at an image, you will always fail to generalize. So what's happening is that these models are actually trying to remember some things in the particular images they've seen and trying to understand. But the real reason why a doctor is calling a chest x-ray having pleural effusion on the right-hand side, that real concept which we understand, is not present in those models. So we see that very obvious pleural effusions, for a model which is trained in North India, are often missed when you showcase a very similar picture having a pleural effusion in South India. And it's just because they have never been trained on those kinds of chest X-rays.

Dr. McDonough:
What's the gap as it exists now between research success and bedside usefulness?

Dr. Datta:
If you look at the research space, the research is happening at a very fast pace today. However, what we have also seen is that the regulations and the best practices of how to use and deploy AI are still not ready.

So what's happening is that people are confused: do I buy this AI tool, and do I allow it to be autonomously used for reporting? Or do I allow it to be autonomously used for just triaging so my doctors can see the patients who are at a high risk first and the low-risk patients later? Or similar for radiologists, if you can see the high-risk scans first and the low-risk scans later.

As the regulations and the rules that tell us how an AI model should be allowed to be used in a setup become more and more clear, I think more people will start adopting, and we'll also be learning a lot more from real-world deployments. Right now, less than one percent of the healthcare institutions which we work with actually are using AI to its fullest capability. So we really are not seeing the real-world failure modes that would happen once you start deploying AI at a significant level. And that is important for us to understand also, especially as long as it doesn't harm our patient, right? If you do it in a parallel kind of way, to see what is happening with the patients when you deploy it in the real world, it will help us understand those failure modes and train the future generation of clinicians with those failure modes so that they would know that, “Okay, even if I am deploying an AI, I should not become too much complacent or too much reliant on its output.”

Let's suppose I see a fracture of a clavicle mentioned in the report. I should not always take it as a fracture of a clavicle because when people have done their deployment research, they have found out that these particular models might be 100 percent correct in diagnosing pneumonia. But whenever it looks at fractures, it often makes mistakes in which particular bone is fractured. So this knowledge, which comes from real-world deployment, is super important. And radiologists have been a little bit fortunate in this space because we have been deploying AI a little bit earlier than the other specialties. So we have seen all of these failure modes coming, and we have been building these processes, which helps us ensure that our physicians and radiologists who are using these AI tools don't get into this complacency loop, start believing all AI outputs as ground truth, and stop looking at those outputs in detail using their clinical reasoning.

So I think that's super important, and we are still missing out, because we have not been able to deploy AI at a level at which the research has progressed. And only once we are able to see the real-world failure modes will we be able to build those processes to make it much safer to independently deploy.

Dr. McDonough:
It's interesting. This is a rudimentary example that's 20 years old, but I know in my own experience we will see EKG interpretations, and you're like, "Wait a minute, that's wrong." Normally they're right, but then you're like, "Wait a minute, that's wrong." I'm looking at this EKG, and the interpretation's not right. At a much more sophisticated level, what you're talking about is a complete review of an X-ray where there might be weaknesses in just one area—like you said, fractures or something else—and the rest could be good. But unless you know that... when you're doing, looking at model performance and clinical readiness, is that the difference you're talking about?

Dr. Datta:
A lot of models behave in very different ways. So today there are almost a hundred chest X-ray AI models, right? Different models have different kinds of failure modes. So some of the models are really bad at looking at or detecting clavicle fractures. Some of the models are really bad at picking up small nodules.

And it has a lot to do with the kind of data it was trained in because, at the end of the day, you have a dataset which helps you build the AI model. If your model has more pneumonia cases, it will often be better at detecting pneumonia. But let's suppose if you look at the distribution and look at the tails of the distribution, the long tail problems which we have in medicine, right? Something which is very rare, like clavicle fracture, or a small lung nodule at the base of the right lung zone. These are the things that often get missed. And this is because in the training data, those models have never seen it. Some models and some vendors might have access to specialized datasets which have more of these fractures and more of these nodules. Those models will be better at picking up nodules and clavicle fractures, but then they will be lacking in the ones which they have not seen in the training data.

So what we are also trying to understand is what kind of failures happen in what kind of models, which is very important for doctors to understand before they start becoming reliant on these AI tools and deploy them in their practice.

Dr. McDonough:
I like the fact you call your lab CRASH Labs, given the different read. But you also have something called RadLE, or Radiology's Last Exam. Let's start simply: what is that?

Dr. Datta:
The RadLE work is our fundamentally most important work we undertook. And this started off, very interestingly, when one of the very well-known AI model vendors was having their launch of a very famous model. And what they did was they brought a patient to the launch of that model, and then the patient said that, "Yeah, I was going to my doctor, and my doctor said that I have two options. But the

doctor was not very sure of what option should be given to me. So I went back and I asked my chatbot, which of the options should I be taking? And then the chatbot was able to answer, and then I told my doctor, 'Let's go ahead and use this option.'"

What happened because of that particular video was that the patient community starts having a level of mistrust in the physician community, and they start thinking, "Okay, maybe the doctor was not sure," which is why they asked us to take the decision ourselves. But that's a choice, right? A management choice which we give to our patients when we have two options. It's not because of a lack of knowledge. But it became very obvious that the patient community was looking at this as an alternative to the physician community.

So as a research group, we were just starting at that time. We felt that it's a very important problem to tackle. Our patients, at the end of the day, go back to their homes. They take a picture of that particular scan they have, and then they ask their chatbots—whatever they use, right? So how do these chatbots really perform for that use case? And what was the use case? That if you give it a single image and ask, "What is the diagnosis?", how good are they, especially when you show them real-world cases? And we also wanted to benchmark them against the radiologists, as well as the radiology trainees who are there.

So we collected a good amount of cases. We tested the four top models that at that time people were using. And then in October 2025, we showed for the first time that there was a decent gap between the radiologists and trainees and the best-performing AI model.

However, very interestingly, in December, when we redid the benchmarking with the top models of December, and Gemini had come out at that time, Gemini suddenly surpassed the radiology trainees for the first time, and these were really hard cases. So we were like wondering, "Okay, so is this how the AI advancement is going to happen? Is this the pace at which AI models are going to become really better than doctors?" So we sat down with clinicians and tried to look at, "Okay, let's have a very independent evaluation of whether these AI models are becoming really good."

What we also realized was that these models very confidently lie to you. Even if they don't know the answer to a particular diagnosis, they will pretend that they know, and then they will say, "Okay, this is the answer." So this is where we realized, as a doctor, if you are unsure of a diagnosis, you can look at the image, you can look at the entire clinical history, and you can still be unsure, and then you will actually go back, discuss the case with your colleagues, look up in the textbooks or the internet, and come back. But the models do not do that. Models inherently tend to lie if they don't know the answer.

So this is what we did, last month when we came up with the second version of RadLE, which is Radiology's Last Exam, in which we gave the models and the human beings an option to say, "I don't know." And if they're giving the answer, we also gave them the option to tell how confident they are on a Likert scale of zero to four. So if you're super confident, you get a score of four. If you are just essentially guessing, you can tell that and we give a score of zero. And then you multiply that by plus one or minus one depending on whether you get it correct or not. Here we found out for the first time that there's a 250 gap between the human radiologists and the best-performing AI model. And most of the models are really bad. And these were really tough cases we showed to them. They confidently said that they knew the answer, but then they confidently made the wrong answer.

So you can imagine that if we really start leaning upon these and deploying these at the patient level, they will never have an idea that the model is telling something completely false to them. They will blindly believe it, and they might not even believe the physician who is telling them something different, or worse, they don't even go to the physician with that particular scan. They think, "Okay, my model is saying I'm all perfect, so I don't need to visit the physician." But the model will miss certain findings. So that is the reason why we were doing this.

Dr. McDonough:
Yeah, if I'm interpreting what you're saying with RadLE, the second version of it, you realize that AI tries to please, and it tries to get to the answer. You're now building it in such a way that it's okay not to know, and since it's okay not to know, you then will get a better picture. It's frightening for me that I probably would turn to AI on the toughest cases when I'm confused, and that would be the one that's most likely not to tell the truth. As you said, there's, thousands of models. Is it possible to get to all of them and do that? I mean, you know from the ones you've tested, but when we get to the point where it's out there, it's happening fast. How do we, one, get the word out, but two, know that it's gone through these steps?

Dr. Datta:
So if you look at the regulatory landscape, ideally the regulatory bodies—FDA in the United States, CE, in Europe, CDSCO in India—they need to lay down the rules of what needs to be the minimum bar an AI model has to pass before it is allowed to do medical diagnosis. Unfortunately, we are not seeing the urgency which the regulatory bodies should show. And probably, the urgency will come, and this is quite unfortunate, once something drastic or something negative happens with a patient who has relied upon this bot and did not go to a doctor.

The only thing which we can do as a community is to bring out these independent studies in front of the regulators so they're aware that these chatbots, which are available to the people and are very much giving medical advice to our patients, are actually very dangerous because they do not tell the patients that they do not know the answer to the question which they are asking, even if they do not know.

Models will always keep on getting better, but it is very stupid of us to assume that models will always be 100 percent correct, right? Medicine, as you and your viewers also know, is a science of exceptions. And every patient we see will never be exactly like the same patient we have seen in our lives. Everyone is very unique. That is something very important for us to understand, which is why it is very important for us to have an accepted bar of safety for the AI models that are allowed to do medical diagnosis. And I think the world will definitely catch up, but we are not sure what it will take for the regulators to understand this urgency and bring these rules in.

Dr. McDonough:
To draw an analogy, which is somewhat frightening, if we look at social media and social media tools, for instance, and how regulations have fallen way behind the expansion and rapid use, you could argue many adolescents are having increased cases of anxiety and issues because of the ways they're using social media and maybe the lack of regulation or concerns around it. I see AI moving even faster, and in medicine, the path has always been slow. We've moved on. Teams have gotten together. They've made rules. They do this and that. Are we able to move at the pace it's expected? What you're doing is clearly showing the issue and getting ahead of it. But will we be able to get to that point, from your perspective? Or how do we achieve that? It's a big question, but what's your thought?

Dr. Datta:
I don't think there is any other alternative. We have to achieve that. What cannot happen is that the pace of advancement of these medical models available to consumers should not outpace the regulations to such an extent that we do not have any way to regulate them later. And I'm an optimist, right? We do work with regulatory bodies here in India as well to help them understand how to do evaluations of models. And I know, there are multiple research groups who are working on evaluations of AI models, and once more and more such research comes out in public where they're able to showcase where these are actually missing out on certain dangerous or critical cases, the regulatory authorities will not have any other option but to intervene.

Dr. McDonough:
Yeah, I think we're seeing this emerge, and I want to get to the clinical application. But before I do though, I'm thinking of the experience probably around the world, but in the United States with regard to COVID, people in general, the population really wanted absolute answers and didn't necessarily understand that science isn't really absolutes. Like we're saying, you're always learning and you're always growing, but they want absolutes. And people hung onto statements that were absolute, and then went back and attacked people because it didn't turn out that way. You mentioned the analogy of a patient who will see the image, and they get an absolute answer because that's what they want from whatever tool they use. But then you get the answer, and the trust is broken. That's another issue. How do we prepare the population to understand that complexity? We spent our careers in medicine, so we get it, but other people just expect yes, no, right, and wrong and don't understand necessarily that it's growth and learning from previous experience.

Dr. Datta:
It's a very tough question, but I think the answer is pretty simple. If you look at the amount of work which the medical fraternity does to share awareness among the medical community is significantly more than what the rational medical fraternity does for sharing awareness among patients. And there will always be one or two people who are probably not so rational but going out there and talking about certain things which have no scientific backing. And they have a huge fan following among the masses and the, and the patient community, right? And that's a typical way how psychology works. You tend to believe the outlier, which seems to resonate more with you than community, which is trying to back science, but you are very skeptical about it because you are not aware of how that kind of science works.

So I think we really need very strong medical influencers who have to go out there and talk about these with the public and genuinely try to tell them, “It's not that we are trying to save our practice, but we are doing this for you and for your health.” The general notion which the public has that healthcare is costly everywhere, all across the world, it really affects the patients when you have to bear a healthcare cost. And you are extremely insurance-driven in the United States, but in India, it's mostly out-of-pocket expenditures which our patients do. So to them, healthcare, and going to a doctor or a physician doing the tests, it's looked at as something which takes away a lot of capital and is very capital intensive.

And on the other hand, if somebody comes and tells you that there is a free chatbot where you can just ask whatever questions you want to ask, and you don't need to pay anything, and you don't need to do a test unless the chatbot tells you, you would really be attracted or, be incentivized, in your mind to believe the people who are reinforcing that kind of a thought, rather than the people who are trying show you evidence why it doesn't work. But that's how the mind works.

So I think the medical influencers who are out there, the rational ones, we need to really back those people and help them spread awareness among the public about these dangerous, safety issues with the medical chatbots, which are all out there in the open to be used by the consumers, or our patients, as I may call them.

Dr. McDonough:
So I know what a lot of our listeners are thinking right now as they're driving down the highway or heading to work or going on vacation listening to this. Wherever they are, they're going, "Alright, I'm going back to work. I'm a radiologist. I'm a clinician working with a radiologist, whatever I'm doing. What does a health system or a physician practice need to consider before using an AI imaging tool?" What's the first question you even ask?

Dr. Datta:
I think the first question we always ask is, "Why do you want to use AI?" There are multiple reasons. So if you look at India, India has a very significant backlog of unreported scans. In India, a resident would have to report 100 to 200 scans every day, while if you look at Japan, Japan doesn't have a problem of too many scans. For them, they want to not miss any findings. So in Japan, the kind of AI models which are actually being used are the ones which look at the particular scan after a human radiologist has written the report, and then it tells them, "Can you take a look at this particular area to see if there is a nodule? Because I think there is a nodule." In India, it's a very different use case. In India, we want the AI to create the preliminary draft for us, and then we go ahead and make the changes in that draft, and then sign it off.

Initially, it looks very simple. But as you start using these tools, you get into something called automation bias and complacency bias. So you start believing that these tools are really good, so you stop checking. So in India, what actually would happen is that the residents who are using AI tools will save a lot of their time. They might be able to see even more patients with that time is the same, right? And they are going to believe and stop editing the AI outputs after some point of time and miss a lot of findings. In Japan, the same thing might happen. They always know that there is a really strong AI model that is there to pick up the findings you might miss. So in the reports that you make, you gradually start being a little bit more casual, and then the AI might miss the same finding which you are also missing.

So it totally depends on why you want to use AI. You have to see, what is that use case which is fit for your particular setup? In the US, I know how things started. People started building different models, and then they realized, "The best possible use for my setup probably is to triage. Instead of telling me the diagnosis, I want to triage that the high-risk scans I want to report first." So the AI does that triaging and puts the higher risk scans at the top of your work list and the lower risk scans at the lower part of your work list. That is how the AI adoption in US started. That is how FDA also started clearing the algorithms just for triaging.

Now, as we are seeing, the FDA is also starting to give clearances to algorithms which can create reports, which can do the diagnosis, right? And these are only going to increase. So as a practice, it's up to you to decide, "If I want to use an AI model, am I doing it to see more patients and increase my revenue? Am I doing it to decrease the workload because it's already too much?" And then your answer will depend on the challenge you are facing and wanting to solve with AI, because the answers will be very different for the two different challenges we usually use AI for.

Dr. McDonough:
Interesting how different cultures have done things different ways in different parts of the world. And I'm optimistic like you, but I also have enough gray hair to be a little cynical. And when I think of, let's say, for instance, the United States, most physicians—I think it's now 80 percent of physicians—now work in a hospital system. Years ago, they were independent, but now, they've been purchased and they work for hospital systems. So if I'm a company that's developing an AI tool for radiology, I've got to sell it to the C-suite. I'm not selling it to the doctors. That's what I saw happening over and over again with electronic medical records. They're selling it to the C-suite. The cynic in me says, "Yes, they all have great ideas. They want to improve care. But they're looking at the bottom line. They're looking at their own bonuses at the end of the year. They're looking at everything, and they're going, 'Give me something that's going to increase productivity.'" By increasing productivity, that means, "Give me something that my clinicians will use and trust to get more done in less time. And we'll pay them more, by the way, because they're getting more done in less time." Or "We'll tell them they've got to do more to get the same money they got before because we gave them this tool." And yet what could happen is the reverse of what we're trying to achieve as clinicians. But with everybody working under hospital systems and others, how do you control that too? That's another barrier?

Dr. Datta:
The real discussion which happens behind the closed doors in the boardroom is that you are only worried about your top line, bottom line, and how you increase your revenue per patient, right? And there are two ways of doing that. One is to charge your patient more and charge the insurance companies more, or you basically have to decrease the cost. And one way of doing that is to decrease the

amount of compensation that goes to the clinicians and healthcare workers who work in the system, right?

So I really do not know how this is going to play out, but there is something called AI deflation, which we are seeing. And there's a very big school of thought that strongly believes that AI is probably not going to take away jobs, but it is going to make a lot of the verifiable and automatable jobs less attractive by bringing down the compensation of the people. How I see this playing out—I do not know how it's going to play out in the United States because I know about the different rules which you have. But in India, what we are gradually trying to see and also promote is that you should run your own clinic and practice so that you know how you want to increase your productivity, and it does not really it doesn't cost your own practice and life. But if you look at the systems, always CEOs and CXOs will want to improve the amount of money they can make per patient. And unfortunately, AI costs money, so that money they will try to reduce from somewhere else, which has been their major area of expenditure.

Dr. McDonough:
One of the theories that provides some optimism that I have seen is that many providers have gone into health systems because they just couldn't do it on their own with all the burden of EMR and all the extra paperwork, they left practices. They joined because they said, "You know what? Let them do it for me, and I'll do that." The hope, I think, is that AI could take care of a lot of that administrative burden as well and take care of a lot of the things that were, like a DAX copilot that allows you to do notes easier. Those things will take time and give them back to you. People may return to, as you say, an independent practice where then they're making the decisions. And again, there are great CEOs and leaders who are committed. But the odds are if you've got someone who's into healthcare and they spend their life, they're going to try to do what they can to balance profit with care and do that and much more. And I guess the better way to say it is, they know the pitfalls of just looking at profit and not care.

So when you're doing this work, are these questions that come up in practice? Are these things that you guys are talking about? I'm sure this isn't the first time you've heard these points. Is that something that you try to address?

Dr. Datta:
Oh, this is definitely not the first time I heard this because this is something that is happening everywhere, right? The problem is definitely there in the United States, but this is a very similar problem we have in India as well. In fact, a lot of the promise AI is bringing today to the independent practice is that it takes care to a huge extent of the different paperwork, admin work, follow-up of your patients, and all the revenue leakage which happens.

I advise a lot of startups. In fact, one of them is in the US, which is helping independent practices cut down the administrative costs. What we have seen is that the major amount of workload reduction or efficiency improvement, in the United States and in India, comes from AI taking care of the administrative burden. And I think this is the time when we need to really use this capability for our advantage so that we can continue to thrive in our practice.

And of course, I know that most of the doctors who have gone through that residency training really care about the patients, right? So I personally feel that independent practices armed with AI for the administrative part would be able to spend most of their time for the patients thinking about the diagnosis, of course using AI tools down the line to help them come up with more differentials and other tests which they might not have thought of. But definitely, I am personally a very big believer that independent practices are going to benefit significantly by adopting AI in their practice.

Dr. McDonough:
Yeah, I think you're right, and I think the other thing before we close this section of it, is that if AI is doing the paperwork and it might not be perfect, I'm not that concerned about the paperwork as I would be concerned about the clinical. So as it's learning and it's developing and as we're testing things out, you can probably move that part forward a little faster than some of the clinical things.

So let's move on. There's a term thrown around called autonomous systems, and when we look to the future, there's a lot of talk about autonomy and what it could look like. First of all, can you define it for those listening? What are we talking about when we say autonomous systems and autonomy?

Dr. Datta:
So I'll tell you how it came out. Especially if you look at LMICs, or low- and middle-income countries, the healthcare penetration of quality doctors is very low. We have around 19 radiologists for one million people in India. And if you remove the radiologists who are reporting for the United States practices, like doing the pre-reads or who have retired, there are just 15 radiologists for one million people. And this is a very similar number for many specialists in India, Africa, and a lot of the Global South.

Now, the only way we are able to provide quality healthcare to billions of people without access to specialists who typically tend to concentrate in the urban areas is through AI chatbots. We have to figure out a way in which we can deploy safe chatbots to the entire

population who can do the initial triaging and ensure that if a patient is having some symptom, they don't go to a local, non-trained person, take their advice, and have some herbal medicines, but rather go to a specialist which is available to them. And we can only do that if we are able to provide them with that kind of knowledge in their own language that they want from a person. If you look five years back, we would not have been able to do that. If you look at today, we are able to develop AI chatbots in all languages spoken, even in a country like India, which has 20 plus languages and more than 50 dialects.

Now, when you have to reach out to these people, it's almost impossible to always be sure about what information these chatbots are giving. And here, these chatbots have to be given some amount of autonomy for decision-making. As regulators and people looking at the compliances, we have to figure out, what are the things which an AI model should be given reasonable autonomy for doing? And what are the things where AI model outputs always need to be checked by a physician? And finally, what are the things which need to be only done by a clinician because AI makes mistakes? It's not that great, and we don't want physicians to get into complacency or automation bias.

So this is exactly what the entire research field of autonomous agents is looking at today: how do we evaluate autonomous agents? What are the use cases where we can allow autonomous agents to be deployed? And this will actually depend a lot on the existence of robust health systems. If it's very robust, like in the United States, you might not need an autonomous bot. But if it's something like a village in India where we don't even have a single trained doctor for that village, we would rather give them an autonomous chatbot with the knowledge of medicine than just asking them to go to some pharmaceutical person who can give them over-the-counter drugs for their symptoms, and then they end up not getting into the healthcare system at all or come to us at a much later stage, let's say, stage four cancer, which we could have picked up if they came earlier to us.

Dr. McDonough:
I think you're establishing and explaining quite well the need, and here in the United States, I see very similar things. There's pockets where you can't get care. Even telehealth has started to reach people, but now we have greater opportunities, and you're right—if somebody gets an answer, it may not be perfect, but perfect can be the enemy of good, too.

Okay, so what kinds of tasks should remain human-led? What would you say are the things that we can't just turn that over, even in the toughest cases?

Dr. Datta:
So there are different kinds of human intervention when you look at AI. The most popular intervention is something called “human-in-the-loop,” which means that all the different AI model outputs are checked by a human. When you look at the diagnostic abilities of AI models, it has the world's knowledge, but it does not have the local context. You might know a certain fact about a patient which the AI model will not know. So using that local context to give the final diagnosis and management for that patient is something which human beings have to do. AI models cannot do and probably will not be able to do for a reasonable amount of time without that context. You might know that the best treatment for this patient is a 30-day regimen, right? But you also know that your patient is very unlikely to follow that regimen. So we would rather give him probably a more costly one-week regimen instead of asking him to go for a much more affordable 30-day regimen. But your AI model will not be able to know that.

So this is a very simple context I wanted to bring to you because our reasons are very different. For us, there are very tribal practices in some villages, like putting oil in the nose, which causes certain kinds of diseases in those particular kids, right? The pneumonia which they get is very different from the other pneumonia. So only the local people know about those practices. So this local context, wherever it’s involved, will need to have humans in the loop.

Some things where context does not matter are triaging of scans, deciding of giving of appointments, prioritizing them. These are some things where we can still do human out of the loop, which means two things. One is a human on the loop in which the models take care of maximum amount of work, but when they are uncertain, they ping the human and say, "Okay, this particular case looks a bit odd to me. Can you please take a look?" What are the kinds of things? They can be follow-ups. That's an example of a human out of the loop in which you are allowing the AI model to make continuously updated decisions.

Out of 1,000 cases, even if the AI makes a mistake for one person, that's harmful for that person. So even 99.9 percent accuracy is not really apt for human out of the loop. So these kinds of things I personally believe always have to be shown to a human being. But some things like which specialist to go to after listening to your symptoms, you have to decide, should I go to doctor A, who is a cardiologist, or should I go to a doctor B, who is, let's say, a gastroenterologist. Maybe the patient has some heart pain, but it is very obvious it's a GERD symptom. So these models can actually tell them correctly, maybe go to a gastroenterologist. No need to go to a cardiologist. So these are the things where we feel we can still give some kind of autonomy to the patients, because we know at the end of the day, they are going to visit a doctor.

Dr. McDonough:
I have a million questions, but I've got to wrap this up for you too, so I'm going to ask you two more questions. For a practicing physician listening tomorrow or next week after we've done this conversation, and they're listening to us, what is one question they should ask before they trust an AI tool? Just one. What would you say they should ask?

Dr. Datta:
I think the most important thing is, have you done external multicenter validation of your AI tool? If you want to use an AI tool which is going to help you make diagnostic decisions, that is super important. Alongside that, and this is probably the case if you are trying to use an AI software or tool for automating a lot of your practice, which is like administrative AI, you will really need to ask, how much time did you save in the previous center where you were deployed? How much was the revenue saved? How much was the efficiency gain? And most importantly, what was the satisfaction of the patients through your system? Because your agents will now be handling appointments and answering the calls. So I really want to know, will my patient be satisfied? Will I really be able to improve my efficiency? Can I increase my revenue? And so it totally depends upon the kind of AI tool you are trying to adopt. But the most important thing is that you have to define the KPI, or the key performance indicator, for you to be able to make an investment in an AI software.

Dr. McDonough:
And my other question is, what is the one thing you wish every physician understood about responsible AI evaluation?

Dr. Datta:
I think, the most important aspect of evaluations which people don't get is that you will always see—and this will become increasingly common—people come to you and say, "On that leaderboard of clinical decision support, my AI tool scores 90 percent accuracy, or 3 percent error rate. We are the best." Benchmark or leaderboard metrics almost always never work out in the real world. So you must do a local deployment of those AI tools before signing a large contract with them to be able to be sure that the particular vendor which is coming to your practice with certain metrics is based on a leaderboard. It can be a very famous leaderboard by Stanford or Harvard, but still it will not matter if that is not the best model for your practice. And we are seeing it very commonly. There are AI vendors which make the tools for Tamil Nadu, which is a southern Indian state, and works very well there. It's great on leaderboards. But then they don't really work well when we go to North India, right? And the same model which ranks fourth or fifth on the leaderboard works very well in North India compared to the Tamil Nadu leaderboard.

So I think all models will have their benefits, the pros, and they'll have their cons. You need to understand which is the best model for your context and your local deployment, and that you can only do when you do a local on-site validation initially before signing the larger contract.

And that is something you must do for any AI tool which you use.

Dr. McDonough:
Dr. Datta, this has been a really useful way to think about medical AI. I want to thank you so much for joining us on *The Convergence*. I really appreciate it.

Dr. Datta:
Thank you, Brian. It was lovely talking to you.

Dr. McDonough:
For ReachMD, I'm Dr. Brian McDonough. To access this and any other episode in our series, visit *The Convergence* on ReachMD.com, where you can Be Part of the Knowledge. Thanks for listening.