(PLEASE NOTE: Transcribed automatically by Vimeo, mistakes are possible/likely. Our apologies.)

TRANSCRIPT - GR 10 27 23 "AI and Clinical Decision Support" guest speaker Ravi Parikh MD from University of Pennsylvania

---

Medicine Grand Rounds

Alright good to see you guys, and good to see everyone virtually on zoom welcome to medicine grand rounds. I'm so excited to introduce you to Dr. Ravi Parique. Dr. Perik is a graduate of Harvard Medical School, Harvard College and John F. Kennedy, School of Government. He completed a Residency in Internal Medicine at Brigham and Women's Hospital, and a Fellowship in Hematology Oncology at the University of Pennsylvania.

He has received the National Palliative Care Research Center, cornfield scholars, award conquer Cancer Foundation, young Investigator, Award, and the A. MA. Foundation excellence in medicine leadership award. Doctor Parik is an assistant professor in the Department of Medical Ethics and Health Policy, and the Department of Medicine at the University of Pennsylvania. He's associate director of the Penn Penn Center for cancer care, innovation and director of the Human Algorithm Collaboration lab. His expertise lies in delivery System reform and informatics. His work is focused on 3 core areas. the use of health technology to improve routine, patient care quality of life and survivorship and oncology and payment reform for advanced illnesses. His work on medical technology and Advanced illness is featured in leading academic journals. Like Science, Nedgem and Jama. He is senior clinical advisor at the Coalition to transform advance care, and was elected the National Council of Resident Fellow, members of Acp. He is here today to discuss with us the intersection with AI and medicine. Please welcome me in joining, or please join me in welcoming Dr. Ravi. Paris.

Ravi Parikh

00:19:13

Thanks, everybody. Thanks, Kara.

- It's a pleasure to be here virtually sorry I couldn't make it to Charlottesville in person getting caught in between travel and some family related things. But I hope to have a little bit of an interactive discussion today. And I've titled the talk. How can clinicians trust AI.
- And yeah, you can sort of read that in 2 ways. You could read it like sort of skeptically like, how can clinicians trust AI, or you could sort of view it as a process like, what can we do to make clinicians trust some of the AI innovations that have been out there, but that we don't currently use in practice, and I'm hopefully sort of gonna address both here. But in addition to what Kara said, you know, I'm a practicing oncologist. I practice here at Penn and at the Philadelphia Va. Medical Center. And I run a lab that essentially tries to answer the question, what types of AI. Can we trust to use routinely in clinical decision support?
- We, you know, answered those kind of questions by running mix methods, studies of among clinicians to assess how they would prefer to receive AI related output, if at all.

- And then we run prospective trials to test whether you know AI tools integrated in clinician workflows can actually improve certain outcomes for both clinicians and for patients. And so, you know, one goal I'd love to have after even after the talk. You see my email there, feel free to email me, if there's any implementations that you've been thinking about would love to collaborate. We're running some multi-center studies now. And we have a really keen interest in answering questions that are important to to docs in across variety of specialties.
- So you know, I think maybe to motivate the talk a little bit, I'll just chat a little bit well, so first off we can go over an agenda here. I'm gonna first talk a little bit about examples where AI based clinical decision support is failed. I'll go over some examples, including examples from our lab where AI based clinical decision support has succeeded, and then I'll try to end by discussing some thoughts and some evidence on how we can make AI in addition to being more accurate, more trustworthy to clinicians.
- So maybe the motivate the talk a little bit. I'll just talk about my first example that I know of being exposed to AI based clinical decision support. So you know, I was second year resident at Brigglin Women's in Boston. We had just roll over to epic and when I came in, my first rotation as a second year was on the inpatient oncology wards and I came in looking at a dashboard like this. And so this wasn't exactly it, but it had sort of everyone's name and room number and reason for admission, and it looked very clean compared to the prior, you know, dos-based, you know. VR, that we had had. That was just super clunky. So I was pretty impressed by all of this, but at the end of the dashboard was a column, kind of like what you see here. It was titled Readmission risk and what you could see was that next to every patient was a red, yellow, or green, indicating whether they were at high red, intermediate yellow, or low green risk of being readmitted. And so they had like a little information link, or you could click somewhere, and you could look into what the score was. And it was actually a sort of your bare bones machine learning algorithm that had been trained on a variety of hospitalizations, and had actually had pretty good performance metrics as measured by common statistical metrics like your positive predictive value, or your C statistic so I was pretty impressed by all this. And so, you know, when we're interns and residents, you know, we organize all the vital signs and everything, and we present to the you know, the dock on rounds, and I think, even though second year resident, I was presenting to the attending that day for this patient.
- And so I said, Okay, you know, this is Mr. Jones. And he was going over his vital signs. And you know, most oncologists, when you like, go over the initial vital signs and labs that kind of like on their phone and are looking forward. But then I mentioned that. You know, this guy has a 80% chance of being readmitted.
- And so the oncologist that was the first time he'd ever heard of that. So you know, he took office classes. He's kind of like. Oh, what? I said. No, he has an 80% chance of readmission. And he said, Well, how do you know that? And I pointed to this score? I said, look at the score! And so we looked at the score, and when we looked at the validation and everything like that. And the oncologist the attending oncologist said, Well,
- I bet everyone on this floor has an 80% chance of readmission. Because, you know, they're all oncology patients. And we looked. And sure enough, you know, everyone had an 80% chance of admission. Everyone was red.
- And so if we you know you know, the oncologist is like, Well, what am I supposed to do with this? Because, if I just listen to this score. I would never discharge anyone for the oncology unit, and my hospital administrator would get mad at me.
- And it, I think is a perfect illustration of you know some of the ways why AI hasn't really made its way into clinical care and clinical decision. Support routinely is because so far we

focused on getting good predictions and putting those numbers in front of clinicians without really understanding how those clinicians, how those outputs ought to be structured and what you ought to do in response to those outputs.

- And so not just me, not this and this oncologist. Everyone on the floor was sort of annoyed by this risk score that kept on popping up and sort of arguing against discharging patients that we really wanted to discharge. And so the good thing about epic is that you can actually remove columns pretty easily from dashboards, and by the end of the first week everyone had removed this thing from their dashboard because it was just more annoying than helpful.
- Okay, a second story. So this was a trial that just came out last week. Model guided decision-making for thromboflaxis and prophylaxis and hospital-acquired Thrombo. Thrombolic events among hospitalized children and adolescents. This is published in gemma network open and basically it was an intervention from the pediatric literature that tried to risk stratify patients for prophylactic anticoagulation if they were hospitalized and so they had developed and previously published on this prediction model that was used to estimate the risk of developing a blood clot. So in the usual care arm, the model predicted risk for everyone, but no action was taken based on that risk. Everyone was blinded to the risk score. and in the intervention arm if you were low risk, no action was taken. But if you were high risk, the hematology team prophylactically reached out to the primary care primary team on the hospital service and discussed whether anticoagulation was warranted. So it seems like a pretty you know, interesting and unique trial and example of where we ought to be going with some of this kind of stuff? So here are the primary results.
- And so what they showed. If you look at the anticoagulation rate you can see that in the intervention arm. The overall rates of being anticoagulate looked a bit higher than the rates in the control arm. But then pay attention to the y-axis. These aren't big differences. These are like sub percent at one percentage point or around one percentage point differences here. And you would expect that in the intervention group. If the hematology team is talking to the team about an coagulation, it would be based on a you. You'd have higher rates of anticoagulation. You didn't really see much there, and as expected, if there's not much higher rates of anticoagulation, you might not expect that much difference in blood clot rate, and in essence you didn't, you didn't. Even though this was a prophylactic intervention. You didn't see any difference in blood clot rate here. And so there was a lot of discussion on Twitter or ex and I'm borrowing these slides from F. Perry Wilson, who I wanna credit because I think he encapsulated perfectly. Why, interventions like this, even though they're based on a on a really potentially sophisticated AI algorithm might go wrong.
- And so when you have. When you delve into the consort diagram, you see a little bit some reasons why the intervention may not have succeeded, and just a little foreshadowing here. It doesn't have much to do with the algorithm. It has more to do with how it's deployed. So kids in the intervention arm that develop blood clots. You know, among those that calculate develop blood clots. The model only didn't calculate as high risk meaning that model only under identified 6 of them. So that means that the true positive rate, the sensitivity of this algorithm was pretty good among the 71 kids that had blood clots and were flagged by the model. Here's some point, some insights into why this intervention didn't work so only 16 were recommended to be anticoagulated by the chematology team. So that's a huge drop off between what was flagged and what was actually recommended to be anti coagulated. So there's some element of trust that we're missing here, that results in basically much less than a third of or around a third of patients, or less than a

third of patients ultimately being anti coagulated, even though they were flagged by the model and among those that were recommended to be anticoagulated by the hyper hematology team. Only 7 were, and this gets even further into the dissemination Gap, whereby you know, the hematology team recommends anticoagulation to the primary team. And there's a lot of. And then the primary team doesn't anticoagulate. That might be for a variety of reasons. Maybe that primary team doesn't trust the hematology team's recommendation. Maybe the Pr. The patient isn't willing to be prophylactically anticoagulated, based on an algorithm based intervention. We don't really know. But there's a huge drop off there. The other One of the big. Other reasons was that. There these alerts occurred on the weekends when there wasn't usually an in-house hematology service that was staffing the model outputs. And so around 40% of the overall, the overall number of kids that were flagged ended up being flagged over the weekend when there wasn't really and a way to act on that intervention. So those people didn't get acted on. And then, for 40 of kids there was a contraindication to anticoagulation for some reason or another that was documented So I think this, you know, also sheds some light into how even the best algorithms, the algorithm. And this algorithm, by the way, had pretty good performance characteristics. Even the best intentioned algorithm oftentimes fails when it comes to scale.

- So I think these 2 examples, they sort of boil down to this overall point. About how we oftentimes approach a lot of the innovation that we see with AI and a lot of the exciting performance improvements that we see in these nature papers and these science papers that demonstrate that AI can do this, and AI can do that. And we judge a lot of these pretending like the AI is going to be autonomous like it's going to function independently in making a diagnosis or making a prognosis because for an autonomous AI and AI that performed independently, really, the main thing that we want to ensure is accuracy. Is it better than a human and does it get it right most of the time? So there's some examples of autonomous AI that's out there, for example, you know a lot of the smartwatch and smart Ekg technologies. They may use convolutional neural networks to arrive at a type of heart room rhythm, abnormality, or arrhythmia that gets automatically flagged to a cardiology consultant or an electrophysiologist. You're bypassing the need for presentation to your primary care to an er and so those might be examples of autonomous AI. The thing is examples of autonomous AI are really low in clinical practice. Most AI that's been deployed is assisted, meaning that it's meant to flag information or flag a risk score to a clinician. But it's the clinician who ultimately makes a decision about whether something is done or not.
- And you know, for on one hand, this distinction between assistive and autonomous is actually really important because the FDA regulates assistive AI algorithms less stringently than it regulates autonomous AI algorithms oftentimes assistive. AI can be cleared by the FDA without a randomized, controlled trial, or even a prospective trial. Oftentimes these things are regulated and approved, based on statistical metrics rather than clinical metrics and so the thing is, if you are assisting a clinician what really matters for there? Well, I would argue that what matters even more than accuracy, even if you had the most accurate 100% oracle like, you know, assistive AI. What matters more is trust, whether the clinician believes in that AI enough to act on it, because ultimately, that's the only way that the AI is going to lead to improved, patient outcomes. And I would argue that we haven't been focusing on this element of trust nearly enough as we should be in our lab recognizing through a lot of failures, deploying AI in actually generating, you know, results that matter to patients. We've tried to adopt elements of trustworthy AI and all the interventions that we develop. And you can see, this is sort of coming from a six-part Figure of trustworthy AI that has been, you know, studied across all sorts of AI. Not just

healthcare. AI, but I think it's it sheds light on elements that often times are ignored. I mean, I'll go through each of these, and I'll go through an example of how we've tried to address each of these in in a particular case. Example? But you know, let's just go through each of these really quickly. So at starting at the top. Fair and impartial meaning is the algorithm not under not systematically mischaracterizing risk or misdiagnosing individuals for a certain sub group of individuals? Is it fair across all the living so subgroups?

- This is oftentimes sort of termed algorithmic bias. When an algorithm is unfair. is the algorithm that you're working with robust and reliable? Does it perform well across different settings, settings, even settings that are different than the setting that it was trained in privacy is the outputs of the algorithm constrained to the healthcare setting that it's being used for? Or could you theoretically be sharing information that goes out into the ether. This is a common concern for using things like Chatgt for healthcare decision making is you don't really know where all that information is going, and for that reason institutions like ours at Penn have actually stipulated that we shouldn't be using large language models for clinical decision making safe and secure are the outputs that are given from the algorithm are the inputs that they're based on protected within firewalled settings, such that they're not accessible to malicious actors.
- Responsibility and accountability. Who's responsible for governing the performance of the algorithm? And is there sort of a point person that you can point to for the algorithm that you're using in case you don't like it, or in case it gets it wrong and transparency and explainability, do you know what goes into the algorithm. And can you explain why the algorithm is generating an output that that's being generated? These are all components that now we started to treat as sort of a checklist for any type of AI or even some sort of stupider models algorithm predictive algorithms that we're thinking about deploying the clinical settings or testing out in randomized trials because we feel like they should be checking most, if not all, of these boxes if we're going to use them for clinical decision support. So I'm gonna move on to a case example that comes out from our lab of where I think we've hit some of these an angles. But we would love to have some discussion if you feel like we could do better. And I don't use this to say that this is the perfect, really great example of how AI should be deployed in a clinical setting. But more so an example where we've used from previous failures to make hopefully a better product.
- And so the use case here is serious illness, communication, and oncology. Now we know from a bunch of randomized trials, and from just clinical intuition that earlier conversations about goals of care and in some cases end of life preferences is beneficial for patients. It's beneficial for patients and caregivers in terms of preparing them for what's coming next. And it's also beneficial in reducing unwarranted care near the end of life.
- Now, one of the key challenges. If you read into qualitative studies around serious illness. Is that you know, time, of course, is a barrier, but one of the big barriers in a busy oncology practice is identifying who ought to have the conversation today. And guidance, like all stage 4 patients, ought to have a conversation that doesn't really help someone like me in clinic where I'm seeing mostly stage 4 patients. How do I know who to have the conversation with today?
- Oftentimes we base these examples on when people have a bad scan result or when they're nearing the end of the line of their sort of guideline approved therapies. But many would argue, that's too late to be having these conversations, and we should be having them earlier.
- So this is actually an interesting machine learning problem. Because if we could identify individuals who may pass away or have a high risk of passing away in a certain time period, then, perhaps, that we could, we could direct conversations earlier to those

individuals than they normally would happen. And that was sort of the use case that we studied in this particular example, that we called conversation connect

- so the first thing that we did in this in this study was, we did some qualitative interviews among oncologists to identify. Is there any role for an algorithm being deployed in your routine practice? Once we found some positive headwinds and saying, Hey, I might be able to use prognostic algorithms to direct things like advanced care planning. Only then did we proceed to actually building a model. And you can see here some publications that we came out with that were around not only the qualitative study, but some of the algorithmic development where we took retrospective data from a cohort of patients that were seen in our cancer center and predicted 6 month mortality among the cohort that was being seen. We then generated a system where we could generate real time estimates of 6 month mortality based on some of the structured data that you see in the box there to prospectively validate an algorithm in real time. We basically ran this algorithm silently in our electronic health record to flag high-risk from low-risk patients.
- And you can see from the graph on the left that we had a pretty good differentiator of mortality risk from our prospective study. You can see that among high-risk individuals they generally had about a 50% chance of 6 month mortality indicated in the blue, whereas among low-risk individuals, they had about a 2% chance of dying in the in the next 6 months. So you know, you can imagine that if you had this information in practice, maybe you would treat the high-risk individuals differently than you would treat the low-risk individuals, and that was sort of the intervention that we wanted to bring to this.
- So, recognizing that it takes more than the algorithm. We then designed a behavioral, economics-based intervention to try to flag high-risk individuals to clinicians, to encourage greater rates of serious illness conversations. So there was sort of a five-part kind of kitchen sink intervention that we developed here.
- We first sent an email weekly that came from our chair of oncology, that detailed peer comparison, detailing how many similar clinicians in their disease group were in their practice, had documented more conversations during that time period.
- That also was accompanied by a performance report detailing how many conversations that individual clinician had documented at the bottom of that email was a link to a high risk list that they could click on to see which patients in their forthcoming weeks panel were high risk. Did they flag above a certain risk threshold that we had previously validated, and we only flagged the top 6 patients. You know, for the for that clinician. We didn't have it. If you were a pancreatic cancer physician, and most of your patients had a greater than 10% risk of mortality. We only concentrated on the sixth highest risk individual.
- There was then an indication of whether they had had a conversation documented or not. Indicated in the sicp. Yeah, no indicator, you see, there and then there was a pre commitment vehicle where whereby clinicians could pre commit to having a conversation for a given high risk, individual and then, if they pre committed to that. Then they were sent a default text message on the morning of clinic.
- indicating that this patient may be appropriate for a conversation. So this is an example of an assistive AI device being embedded in a holistic care intervention to try to boost rates of conversations and we generally saw you know what we would characterize as a positive study here.
- When the intervention rollout began indicated in the red dotted line, we saw a pretty much rapid increase in rates of conversations that were documented over the next 4 to 6 months, such that you know, at the peak of the intervention indicated in the green line.
- High-risk individuals! The rates of documented in conversations increased by about fivefold compared to baseline among high-risk patients, and they reverted back, but not

completely back to Baseline even during the follow-up period of this intervention, indicating some sort of sustained effect.

- This, translated among people who died to some differences that we saw in end of life aggressiveness of therapy, for example, we generally saw a decrease in the intervention group in rates of chemotherapy in the last 14 days of life compared to the control group that was significant and also associated with cost savings. And, by the way, these are this is actually, there's a 0 missing here. It's around $15,000 saved in the last 6 months of life. Oh, sorry this is savings in the last month of life, which is around $1,500. These were magnified when you look broader out to the last 6 months of life.
- So A pretty positive intervention here when you look at the numbers. Now, I'm going to detail now some examples of how this intervention, even though it may look, you know, positive on the surface, actually had some major trust issues among clinicians, and how we've been trying to address those.
- So the first thing that we did was that we interviewed clinicians that participated in our trial. After we rolled this out and and had gathered all our primary data, we interviewed 25 oncology clinicians, and we asked them what went well and what didn't go well, with this intervention and the biggest facilitators to actually having the conversation.
- I would argue, had nothing to do with the machine learning algorithm.
- They had to do with sort of norming of having earlier conversations that was shared across similar cohort of physicians. They had to do with things like peer comparisons and performance reports, and they had to do with better documentation of conversations in the Ehr prompting conversations that many would argue were already happening to actually be documented so that they could be viewed and shared by other members of the care team and this was published in a paper in the Journal of Palliative Medicine. It says in press here, but this was recently published.
- Some of the barriers to the intervention was actually around the algorithm performance. For example, many noted that there was cancer specific heterogeneity in the algorithm performance, even though we were deploying this across diseases, many of our breast cancer clinicians and he malignancy. Clinicians said, for example, that there were a relatively high number of false positives that decreased their trust in the algorithm overall. and many didn't like the idea of how we delivered the messages in terms of text messages rather than you know through the Ehr, or through something like that. And I thought a quote like this was actually somewhat illustrative of, you know, one of the ways that we went back to the drawing board. So the one of the our he malignancy on colleges said in blood cancers. We do so many scheduled admissions for things like stem cell transplants.
- There's a lot of patients who would have been in the hospital twice. So actually, the algorithm based lists were often inaccurate in terms of who needed to have a discussion because they were flagging people that were hospitalized for scheduled admissions and not because they were sick.
- So you know, really sort of humbling. When we were riding high on what we thought of as a positive trial. And so, you know, I want to kind of walk through these principles of trustworthy AI, because we've really taken this to heart as we've tried to revise this overall process. So one of the primary elements of trustworthy AI is robustness. Is your algorithm relevant in populations that are different in than the population you may have trained on. And so you know, we since you're running this trial, we've since externally validated this algorithm in other contexts. For example, in Dana Farber Cancer center and generated pretty similar performance metrics as the original trial. We also showed in one of our publications that the algorithm that was trained on primarily solid tumor malignancies actually had a pretty high performance across a variety of disease groups, including

disease populations like neuro-oncology and community oncology practices that were underrepresented in our training cohort.

- But in many cases the performance differences actually differ quite a bit. And you might imagine that you know an algorithm that's trained in, you know, the northeast might perform very differently in an inner sanity community in the South. And so you need to sort of think about how you know some of those robustness challenges are are gonna take place with your AI, and oftentimes we've tried to be running algorithms silently in our populations, especially if they're off to shelf algorithms so that we can ensure that the performance that some sort of vendor quotes to us is actually held up at our own institution. We're also sort of trying to take this concept broader. And so this is a concept that we're submitting for a large cooperative oncology group. The swag network. Where we're proposing validating this algorithm as a flag to target patients for palliative care in a primarily community oncology setting so quite different than the setting that we rolled out our initial intervention and even externally validated our algorithm.
- And so all of this is is meant to be proof of concept of robustness, demonstrating that this algorithm based intervention can hold in external care settings.
- What about usefulness and transparency? So I would define transparency as understanding all of the elements that go in to a given model.
- And so there have actually been a lot of attempts to make AI more transparent by coming up with sort of nutrition label type things. And so you can see here, this is a concept that came out of Duke out of Mark Sendec and his group at the Duke University Forge team. And so, you know, they developed a model to predict risk of sepsis among patients who were admitted to the hospital and what they essentially developed was a nutrition label for this algorithm prior to its deployment in the hospital care setting. And so they, you can see sort of what goes into this nutrition label. It details what a summary of what the model is. Is meant to be used for. It details. What are the components of the outcome it was trained on. And the input data that it's based on what are the performance characteristics of the model according to a variety of performance. Characteristics. What are the sort of intended use of the algorithm sort of like an indication on an FDA label?
- And what are some warnings about the label, for example, what settings should it be used in, and what shouldn't it be used in? Maybe it should apply to people on admission, but not necessarily when they're admitted to the Icu, for example, and then it contains some resources to look at. If you want more information.
- So we've been in love with this concept. We think that it sort of serves as a neat tool for clinicians to click on for use of algorithms if they want to know more about it. And so we've been developing these for some of our algorithms that can be sort of linked to in the electronic health record. When you're say, for example, if you receive an algorithm prediction, you can click on an information link and have this pop up in an external link. If you want to read more what about explainability? So our lab, I would define explainability as the ability to understand why an algorithmic prediction is getting the result that is being given to you. Essentially, why isn't? Why did an algorithm come up with the prediction that it gave you?
- And explainability is really tough, because oftentimes our AI tools, they live in a black box, and the way that they arrive at an output is through some sort of complex modeling of nonlinear relationships that isn't easily able to be expressed the way like a regression coefficient would be. And so there's been a lot of methodologic avenues because we really haven't solved the explainability problem.
- You can see here an example. Say you had a deep learning model that was trying to predict the pathology from a chest X-ray without explainability it might give you an overall

prediction or give you an output that says, Hey, this patient has cardiomegaly on the scan, and we're 78% confident. But you don't really understand why that prediction was made. Why is it correct, or why might it be incorrect? And what features of the X-ray contributed to that prediction.

- And so you can see a hypothetical example. This comes from a Nature medicine paper.
- That shows what an explainable, unexplainable output could look like. You know. Say, for example, it's giving some text as to why it's arriving at the cardiomegaly prediction, because the lungs exhibit cardiomy with a large silhouette of the heart.
- Say, it's coming up with some sort of heat map visualization that indicates to areas that contribute to the prediction of cardiomegaly those are all components that I would argue if I saw in an output that might lead me to trust an algorithm rather than if it's just spitting out something that says cardiomy.
- Now, I would argue that this has been tried for radiology based AI before, and we frequently got it wrong. So this comes from a recently published paper. That dealt with a concept card called large multimodal models. And so they fed images. Into this Gpt module and asked it to come up with a radiology report.
- And so you can see the output that it came up with for this sample image here which clearly shows, you know, a lung module indicated by the arrow.
- and you can see the findings that the AI, the Gpt module, came up with this chest. Ct. Demonstrates a nodular opacity in the left upper lobe, measuring approximately 1.3 cm in diameter. The opacity appears to have speculated margins located adjacent to the plural plural. There's no evidence of media, stymal or high lower lymphedenopathy. The findings are concerning for a primary lung malignancy.
- And so I anyone. Who reads this is going to look at this and say, what are we doing? Trusting? AI, because there's a lot of things wrong with what it came out with here first off this nodules in the right. Not it's in. It's not on the left. There's no way to make a call about pathologic add andopathy from a single slice of the Ct image. They're not even looking at a lot of the, you know, sub corinial and other lymph nodes that. You know, play a role and how we determine pathologic. Add anopathy. And it's making a size estimate without having a 3 dimensional understanding of what the nodule actually looks like. That's just the least of which any radials just in the room is probably gonna come up with a million other things here. So I think the key here is that frequently when we ask AI to be explainable.
- But we're just feeding random data into it. Oftentimes the AI is coming up with some hallucinations as to why it thinks it's right. Many of you have seen examples like this from experiments with Chat Gbt. And this is a problem. Because if we put out subpar explanations that's going to lead clinicians to distrust the entire system, even if it gets it right most of the time.
- So what we've been trying to do in our lab and this is the topic of an ro. One we just recently put in was trying to kind of reverse the paradigm of AI of clinical AI rather than sort of feeding in a bunch of data and asking the AI to come up with explanations that we just kind of click. As, right or wrong, we actually feed in rules to the deep learning module, for example. with laboratories. Here's a common set of laboratory threshold for long, normal, low, and high. values for certain laboratory values, like lymphocytes, albumin or urea nitrogen.
- And so when the algorithm when it's fed with these rules, and then is interpreting the raw data that's fed to it. It can generate explanations that adhere to these thresholds that we, as clinicians, commonly deal with because otherwise sometimes these AI modules are coming up with thresholds that don't make any clinical sense to us. It's saying something

like, you know, an album in level of less than 4.5 is a is a poor prognostic factor when we're more used to seeing a threshold of like 3.5 as being the low value.

- So you know, coming up with these sort of what we call domain knowledge based explanations, I would argue, is a better way to build trustworthy AI in the last few minutes before I want to get to questions and discussion in a second. But another big component of trustworthy AI is reliability. Meaning. How can you trust that the AI is performing the same as it does now as it did before? And so we know that what we call with what we call static models, meeting models that are deployed and don't update over time. Generally, the performance of these models decreases over time.
- They can decrease because components of the cohort are changing. Your population is changing. We saw a big decrease in reliability around Covid because people's input variables, for, for example, things like labs or chest X-rays or Ct. Scans. They all fell off a cliff during Covid, because no one was coming into the hospital for a lot of things except for Covid and so we saw a big performance decrement in a lot of predictive models. And I'll show you one example in a second.
- And so it's known that if we kind of proactively refresh models and retrain models, that oftentimes we can maintain performance without drifting.
- Here's an example from the same mortality, risk prediction model. II highlighted to you before we saw that during the Covid period, march to May of 2020, there was a huge drop in the true positive rate or sensitivity of the model. You know, from before to after.
- and a lot of that was driven by changes in the way people were coming in for lab visits in our cancer center. There were less lab visits, and in general there were more abnormal lab visits, because we only encourage people to get labs if they were feeling particularly sick, for example, and so you know the you know, this performance strip has since rectified as people have started using the health system, you know more in in normal fashion. But you know, basically, our model was under identifying individuals during the entirety of the Covid pandemic period. And so we had to retrain the model or come up with different thresholds for it to perform the same way that we set it with another critical, critically important element of trustworthiness is fairness. The ability to not mis estimate risk for subclasses of individuals. There was a a very seminal paper that came from a risk prediction algorithm that was used by a major payer to flag individuals for high risk care, management. And they showed that this algorithm was systematically under identifying black individuals for high-risk interventions compared to white individuals. In fact, you can see on the graph on the left, at the algorithm based threshold indicated in the black dotted line. In general, a black individual needed to have one to 2 more chronic conditions for them to flag in compared to a white individual.
- Now, one of the reasons why this was happening was because the actual model
- wasn't actually predicting how sick someone was. It was predicting how costly someone was because the insurance company had much better data on cost than it did about how sick someone was, and so you know, cost is not a bad proxy for how sick someone is, but it's well known that black individuals at similar levels of sickness cost less because they utilize the healthcare system differently than white individuals in certain care settings like this one. And so what these researchers really dramatically showed was that if you, if you use a different label than cost, if you use actually some metric of how sick somebody is you might result in more than threefold increases in the rates of black individuals qualifying for the program. Compared, if you're using your normal score.
- And so this is a very common problem around AI, especially because oftentimes the AI is predicting what's easy to predict rather than what's right to predict. And so tracking how these might be perpetuating systematic biases that are already discovered in the

healthcare system is really important, because if you're a clinician that treats a primarily underserved population. You might distrust this algorithm a lot more than if you knew that it worked for that population.

- Now, what are some reasons. Well, you know, we've actually been experimenting with this for prognostic models for cancer. And one of the most common reasons that algorithms bake in bias is because they're baking in systematic biases that are occurring in routine practice. So say, for example, you're trying to predict risk. You know, prognosis someone with breast cancer. Certainly someone's genetic and molecular and germ line characteristics are going to be a contributor to someone's underlying risk. But you might get a model that shows that for an African or American woman with breast cancer, someone's risk of mortality appears to be pretty low.
- Now, when you dig underneath the surface, you might look and say, Well, what is it actually showing for genetic testing? And it's actually showing that the genetic testing is missing.
- And so oftentimes what these AI algorithms do will be they compute a value for the genetic testing rather than you know, actually saying that we can't predict this for someone who's missing data. And we know, as you can see up from the table here, that on average African American women are under tested for Germline, or genetic testing. Even after accounting for their underlying risk of having a pathogenic mutation. It's only when you account for the physicians propensity to send for a test that you actually explain most of the reason as to why African Americans are under tested.
- And so if you were trying to develop a model based on someone's genetic test. You might actually, really dramatically, under predict risk, because you're baking in a bias around likelihood of testing among African American women.
- Now, I don't want to paint fairness and bias with too broad a brush. What we've showed in our conversation connect trial is that we? Our intervention actually had a disproportionately more positive effect among not among minority individuals, namely, non-hispanic, black, and primarily Hispanic and Asian populations. And that's predominantly because the algorithm was correctly identifying risk in minority populations in whom rates of conversations were much lower. At baseline you can see the pre intervention. Conversation rates were much lower in our minority groups compared to the non-hispanic white individuals. That's a known phenomenon. And so sometimes these algorithms, even if they are baking in some bias can still rectify biases that are occurring in the underlying care delivery system.
- And that's a positive thing. So you know, we've and this is sort of a conceptual model that's based off a paper that came from Isaac and David Bates and Nature. Mpj. Digital medicine a year ago, you know, in the serious illness, conversation, example. We've really tried to think about how we might integrate trustworthy AI at all elements of the process ranging. If you look on the graph from the right for identification of sic eligible patients like we did in our trial to chat bat based technologies to collect information upstream. Prior to a potential conversation between a doctor and patient to enabling earlier therapeutic conversation and shared decision-making by essentially, you know, pre-filling information about that plan beforehand, using digital health technologies.
- Another big barrier is identifying whether these conversations are done from clinical notes. And so we might be able to use natural language processing type technologies to identify whether a conversation was documented.
- All of these are sort of examples, many of which I haven't touched on, of how I think we can build AI and trustworthy ways in this particular use case that I just talked about of

serious illness communication. But you can probably see a lot of examples for other use cases that are more familiar to your own clinical practice. I'm going to skip this slide

- So maybe the like in the last slide. I'll just sort of highlight that, you know, as we're thinking about building trustworthy AI, we've got to realize that there's more than just data science that contributes to whether clinicians will trust the prediction. You know, there's a lot of components that go into a true human machine collaboration that can be trusted. There's the machine inputs, which is what we generally focus on. But then there's also human inputs realizing that humans have unique insights, clinicians have unique insights to contribute. And we should be. You're trying to understand. Where can clinicians supplement poorly performing algorithms rather than viewing it purely from the other side.
- And then there's contextual inputs, for example, trying to, you know, address things like alarm, fatigue, or inadequate resources to act on a prediction or a one size fits all threshold for flagging, because oftentimes these algorithms are flagged as sort of alerts or bells and whistles or early warning systems. And we've got to realize that if they're deployed in suboptimal ways like that epic readmission risk example. I cited at the beginning of the talk.
- Then, you know, clinicians are just gonna click through those type of alerts the same way that I click through all the Bpas that that come up on my epic. And so we've gotta think about how we can integrate those in in the workflows that can that clinicians will actually react to. And sometimes the answer there is bypassing the clinician entirely and flagging that towards an end user that might trust a prediction more than a clinician with so this is a framework that we've used in a lot of grants that we've put together that are around, you know, building trustworthy AI for clinical decision support this is maybe a slide. I think that you know, might appeal to some of the health system leadership in the room. And that's around. What can we do as a help system to enable readiness for AI especially given some of the concerns around trustworthiness.
- And so there's a couple of things first off. Oftentimes, you know, we get advertised. These off the shelf. AI tools. I know our health system gets pitched a bunch of AI tools every day. Yours probably does, too. And you know, they're quoting performance of their algorithms that are based on really optimized care settings where there's a lot of completeness of data. And so when we try to interrogate our own data whether it would, you know, work well for their particular AI. We oftentimes see that our data is just not fully adequate.
- And so what might be one way to enable writing this. Well, I think getting better, you know, going back to the bones of the problem and getting at more completeness of data streams that would enable the AI to function. In the first place another way is by getting Nim nimbler software, you know, as I mentioned, you know what's really important for a lot of these AI tools isn't whether they're necessarily more or less accurate, but how they're presented to the clinician, and that oftentimes is a barrier in terms of the software that's used to present the Api or the Bpa that comes up to a clinician rather than what actually goes into the algorithm and having more nimbler software to enable better interfaces, I think, is a cre is a key barrier.
- Governance is a big priority. You know, oftentimes we find that health systems, including ours. We don't have a centralized governance structure to dictate. What AI priority should we be seeking out. We don't have a way of dealing or sifting through all these vendor requests that come into our individual departments, and so, having an individual governance structure that can. You know we used to define priorities for I. AI, but also, you know, set standards for how we ought to be monitoring and deploying these AI frameworks in our own health systems, I think, is really important, and that'll be a priority for health systems. You know, this might be formalized under a title like Chief AI Officer, which you

are increasingly seeing at some health systems and then proactive monitoring of AI performance, so that we can determine whether the AI is functioning in the same way that we think it is.

- I would argue that if we can get these elements right, if we can make a if we can have. You know the focus that we've put on accuracy and sort of innovativeness of AI. If we can put that focus on building AI that clinicians can trust. Then we'll have much bigger impact than we've currently had in actually using AI to influence patient outcomes. But we need to get there through a reshuffling on priorities. And hopefully, this this presentation spurs, some food for thought. So thanks very much, and would love to have a little bit of a discussion in the last 10 min or so. I'll stop sharing.

Medicine Grand Rounds

01:07:50

That was excellent. Thank you so much. We have. If a lot of people on zoom today, if anyone wants to drop a question in the chat, please go for it. And then we have a question in the audience here.

- Okay, can you hear me? Okay, that's probably working. Thank you so much. That was very interesting to hear about. Many of the examples you highlighted talk, or were examples of the assisted AI models which seem like they have the potential to add time and mental burden to providers like prompting like risk scores or prompting conversations.
- How do you? Or maybe some of the doctors that you've worked with these models feel about that? Maybe, added burden that they face with these things. And how does that play into the general physician shortage in many fields?

Ravi Parikh

01:08:39

Yeah, that's such an excellent question. Sorry I couldn't see your face, but II I think the the the way that at least we've tried to approach some of these problems is by trying to measure a some of the impact of our intervention, not only by the end result, for example, number of conversations, but also in trying to estimate some sort of time burden saved from summarizing information because one of the positive aspects of these AI tools is that they can be used to summarize information that we would normally be taking time

- to look at and so I'll just give you an example one pilot that we've been working on now is AI based decision support for clinical research coordinators. To help aid in prescreening for patients for clinical trials. Oftentimes our clinical trial prescreening process is measured, is is determined by, you know, clinical research coordinators manually annotating, you know. Dozens or hundreds of notes to try to identify whether someone is particularly eligible for a trial or not.
- And there's many, you know, principles of natural language processing that could be used to identify simple elements that could save time. And so we've we're starting to run a prospective trial where our primary outcome isn't necessarily accuracy. We think that there's not necessarily an accuracy gain to be had from some of these AI tools, although

maybe there is. But the primary outcome is time per chart review or number of chart reviews able to be had to identify an eligible patient for a trial. So I would argue that as process metrics, especially as these AI tools start to be regulated more and have greater needs for phase 3 testing. We should be integrating process outcomes around efficiency in there. The other point I generally make is that oftentimes clinicians aren't the best.

- They're resource, constrained, and often times they're not the best end user or the end output for these AI based decision support tools. We might think that clinicians are the major determinants. Say, for example, of whether someone ought to get a care management program. But in actuality, if you actually did some stakeholder interviews, you'd find that. No, it's actually, you know, social workers or care managers or nurse care coordinators that are the ones that ought to be receiving this output and bypass the physician entirely, and that's not to say that care. Coordinators and social workers aren't also overworked. But sometimes there's generally a little more trust to acting on the algorithm and saving time that is manifested on the on the ancillary or care coordination staff than it is for the physician themselves. And so we ought to be targeting the AI towards the people who's gonna help the most rather than just saying the physicians gotta be the end, all user. And all this very interesting. Thank you.
- Yeah, that was that was very insightful. We're passing it to someone else in the crowd. But I had a question as well. When you're talking about. How you just kind of get emails for these random AI protocols. And you don't really know what to do. And maybe we should have AI Coordinator at institutions. Can you give us just a quick idea of the landscape who is creating most of these AI products? Is it mostly like private industry or a lot of, you know, large health systems guiding this like, what does the landscape look like? Externally validate a solution, for example, or from academics?
- If we would trust those a lot more that, I think, than we trust the vendors in all honesty. II just think that there's a the so much of the volume is coming from private industry, whether those be, you know, large scale organizations like Google or Microsoft, or more commonly more startup sort of under the surface organizations that are looking for health systems to validate their product. And so, you know, that's not to say that those solutions aren't going to be game changers, many of them will be but it's just really difficult to sort of separate the wheat from the chaff there, and usually we're only able to do it after we take a couple of hours of understanding what data needs to go into those models. To actually, understand? Hey, this really just isn't gonna work at our health system or not. Usually, that has to do, because we haven't set up the data feeds in a way that will work for the AI output. Sometimes it's because these organizations there's some malicious intent and wanting to share our de-identified data with other organizations. And that's actually how they make a lot of their money. I'm not actually from the product itself. Some of these people offer their services to us for free. And it's really the data sharing that they care most about. And we generally tend to distrust those a little bit more, because, you know, we wanna be sensitive of our patients. Data. I say, we in a very general term here, like I, you know, I I'm involved in some of these conversations because of my research expertise. But a lot of these are being handled by our chief informatics officer and the like.

Medicine Grand Rounds

01:13:48

Excellent and dr. Greg Madden dropped something in the chat for you.

Ravi Parikh

01:13:52

Yeah. So I'll just read it out. So yeah, you mentioned the importance of explainability with AI and medicine. Have you seen any newer explainable ait? Techniques such as Shapley? Additive explanations trickle down the clinical AI applications? What do you think are the best way to explain clinicians, black box predictions. Okay, this is a great question. This is actually the topic of our ro one that we submitted in in October. So shapley Explanations are an example of what we call a post hoc, explan, a explanatory method, meaning that they're being fed in the data raw. And then they're the algorith the algorithm is coming up with explanations for individual predictions based on individual co based on covariates that are used in that that contribute for a given prediction. So say, for example, there's one particular laboratory value that's particularly associated with mortality, and that's what's most abnormal for this patient. A chapley value might flag that value at a specific threshold.

- We've used Shapley values a lot, and we've generally moved away from them because they tend to come up with gobbledygook for explanations, and the reason is because oftentimes they frame things as correlations rather than causations. For example, you know, you might see that someone's mchc, the mean corpuscular hemoglobin concentration, something that I never read in a Cbc. Really, that's what's flagging is the topmost predictor for mortality, and that's because it's a correlation that the AI has no way to determine whether that's causative or not.
- And it's often citing it at a really weird threshold than I'm used to in clinical practice. So what we've generally found is that wha what we are more positive about are so called neuro symbolic methods that integrate domain knowledge and rules first. So here are the relationships that clinicians will believe. Here are set laboratory thresholds.
- So you feed those into the deep learning model and then have the model extract explanations according to that causative framework, and then it generally tends to come up with explanations that are more believable, and the whole point of explanations is to get people to believe in it, so they'll act on our prediction a. A, and so generally we've
- I haven't seen this used yet in an actual AI deployment, but that's part of our RO. One is to see whether those type of explanations are trusted more by clinicians but we probably only had time for one more. There's one more draft in the chat. I won't attempt to read it for you.

Ravi Parikh

01:16:30

Where Gpt is trained on Wikipedia, and common crawls no supplies. How slobbily it is, and pure natural language processing on their own can be with the lung nodule example. How do we rather train Ml. On longitudinal patient-level records in a world where Ehrs are disjointed across the country which harms the generalizability of any off-the-shelf model.

- Federated learning is 10 to 1220 years away at the best, and Lstm models need massive data to input, should we create policies that stop hospitals from using Cds models created by other sites? This is such a good question. And I think really points to the need for better

governance of some of these structures. So I'll just give you my answer to the last one. I don't think that we ought to be stopping hospitals from using these tools. Because many times, you know, there you might expect, like, say, for example, your tool is based on biomarkers from a next generation. Sequencing panel. For example, in general, the next generation sequencing panel is gonna have similar inputs from platform to platform with some differences. And so you might expect that it's gonna be applicable in your institution. But I would argue that the better policy to create at a hospital level is a policy to externally validate a tool in your health system. Before it's used for deployment, and that might convince others that you know, it's even this off. The shelf tool is able to be used because I totally agree that the better solution is to train on larger and larger data sets so that things can be more generalizable. The problem is getting access to those data sets. And you know, for the variety of reasons that you've dictated here. And so in the interim, I think we need to have policies that at least allow us to test whether these off-the-shelf tools, work or not. I don't think we need to have a, you know. Across the board policy that we shouldn't use them, because otherwise, I think we're going to shut ourselves off to a lot of potentially good tools.

Excellent! Well, thank you so much for having thank you so much for coming, and we enjoyed having you have a great day.

Ravi Parikh

01:18:42

Thanks, everybody.