Dr. Lisa Dixon (00:07):

Welcome to our podcast, Psychiatric Services From Pages to Practice. In this podcast we highlight new research or columns published this month in the Journal Psychiatric Services. I'm Lisa Dixon, editor of Psychiatric Services, and I'm here with podcast editor and my co-host Josh Berezin. Hi, Josh.

Dr. Josh Berezin (00:26):

Hi, Lisa.

Dr. Lisa Dixon (00:27):

Today we're going to talk about a paper that uses machine learning techniques to try to predict outcomes of antidepressant treatment. And our guest is really a wonderful, well-known, very productive researcher and as we learn a little bit also a clinician, Dr. Greg Simon.

Dr. Josh Berezin (00:48):

We are very fortunate today to have Dr. Greg Simon, who is a researcher at Kaiser Permanente Washington Health Research Institute here to talk with us about his and colleague's recent paper predicting outcomes of antidepressant treatment in community practice settings. So Dr. Simon, thanks so much for joining us.

Dr. Gregory Simon (01:04):

You're most welcome, happy to join you.

Dr. Josh Berezin (01:06):

And our listeners will hopefully excuse my hoarse voice, but I think we will get through it. So just by way of introduction, can you tell us a little bit about what your current role is and what your research interests are in general and also in thinking about your work, whether there are any through lines or things that you think of that kind of are like your motivational courses in your research program or things that really kind of drive your work.

Dr. Gregory Simon (01:41):

I'm a psychiatrist by training. I would describe myself as a mental health services researcher, although I still have a part-time clinical practice. I still see patients. I work in Kaiser Permanente Washington Health Research Institute, which is a research institute embedded in Kaiser Permanente Washington, one of the Kaiser Permanente regions. Kaiser's a large integrated health system with several regions across the country. I also lead a network, our mental health research network, which is funded by the National Institute of Mental Health and it's a network of mental health researchers and research centers that are embedded in large health systems. So now includes several regions of Kaiser and other sort of similar health systems.

Dr. Lisa Dixon (02:17):

So I have to jump in for just a second and say this mental health research network is absolutely fantastic, productive, teaches us so much about care. So I just need to be a fan here for a moment.

Dr. Gregory Simon (02:31):

Thanks much and I hope our funders are listening. I would like to think we focus on practical research. We're interested in the intersection or the questions that lie at the intersection of things that the members or patients in our healthcare systems care about, that the clinicians in our healthcare systems care about, and to be honest, the thing that research funders care about because that's often what we depend on. Although this particular study is an interesting one, because this actually was funded by some internal funding from Kaiser Permanente, not by an external funder. And so the story here is an interesting story. Kaiser and our mental health departments across Kaiser have been really focused on implementing what some people would call measurement-based care or one of the other terms sometimes used as feedback informed care. But the idea that we should be systematically measuring the outcomes of mental health treatment, trying to understand when people are improving as expected, when they are not, how we might adjust or alter their treatment.

(03:27):

So this project was motivated by the hope that we would be able to use the information in people's health records to be able to accurately predict who might improve or how much people might improve after they start antidepressant treatment for depression. We were hoping to apply this to be able to give information both to clinicians and patients to say that this person is improving as expected, this person is not improving as expected, maybe we should change their treatment. In a sample of about 2,500 people who were starting antidepressant treatment for depression, we were able to extract from their health records, it was about 300 different variables that had to do with the things most people would expect, their history of prior depression treatment, the other diagnoses they received, things like whether they'd been hospitalized or ever had a suicide attempt, depression scores and the PHQ-9 that had been reported in the past.

(04:17):

And we used a relatively simple actually machine learning method to say, "Can we combine all this information and come up with an accurate prediction of either how much people's depression scores would improve or by using that standard 50% or better improvement, what proportion of people would have that so-called response to treatment?" And what we found was the prediction was actually quite poor. It may have been slightly better than flipping a coin, but not by much.

Dr. Josh Berezin (04:43):

Even with all those variables.

Dr. Gregory Simon (04:44):

Yeah.

Dr. Lisa Dixon (04:45):

Greg, could you just add a little bit about what was the population studied? Where? To whom can we expect this would apply?

Dr. Gregory Simon (04:55):

We use data from two of the regions of Kaiser Permanente, which is a large integrated health system. It serves people who are insured or members of these health plans in various ways. People who are employed and have commercial insurance, but also people insured by Medicare, because they're older or disabled, people on Medicaid, people who have subsidized insurance. So it's a relatively broad cross section. We looked at people who were starting antidepressant treatment including people starting treatment in primary care and those with psychiatrists or specialists, although the majority were from primary care. These were people whose baseline depression scores were mostly in the sort of moderate or high moderate range of depression. So it's a pretty broad and I think reasonably representative cross-section of people starting antidepressant treatment.

Dr. Lisa Dixon (05:38):

And this is in the Pacific Northwest?

Dr. Gregory Simon (05:41):

Yes. This sample was in the Pacific Northwest.

Dr. Josh Berezin (05:44):

What sorts of things were you using to predict the outcome? What was going into your model that would hopefully have said this person will improve, this person might not improve?

Dr. Gregory Simon (05:57):

The information that was available in records, which we have to say is limited. So certainly we would know what people's depression scores were at the time they started treatment and what they had been in the past, if people had received previous treatment, whether that depression scores had been recorded in the past. We knew what other diagnoses were in people's records, whether they had co-occurring anxiety disorder diagnoses, which are relatively common substance use disorder diagnoses. We knew about what kind of treatment they had received in the past, whether they had taken antidepressant medications, steam therapists.

(06:28):

We knew if they had been hospitalized for mental health reasons, if they had emergency department visits, if they had made suicide attempts or had episodes of self-harm in the past, all of these things that were indicated in people's records for up to five years. It is important to say what we did not know. We did not know the circumstances of people's lives. We did not know what might've been happening in their lives at the time they started treatment and how certain sort of stresses or traumatic events might have affected them or might have actually improved and leading to people feeling better. Those things may sometimes be written in the notes of records, but we were using the data that can be readily extracted from records, the sort of codes and numbers and things like that.

Dr. Josh Berezin (07:07):

So you weren't going through and using a language model to look at the free text notes to pull things out. You're looking at somebody clicks on this is their diagnosis, this is when they start medication.

Dr. Gregory Simon (07:20):

Right. Looking at the more sort of personalized information that is in text is certainly something that, to be honest with you, we've started doing and that's where you might find much more useful information about whether people would improve. But once again, I had assumed that we would find that outcomes were much more predictable from what we would call the sort of coded or discreet data that we used and they were not. So now I'm open to being wrong about the fact that some of that other information might not be very accurate or the information might be accurate, it just might not be very predictive.

Dr. Lisa Dixon (07:53):

If you were to do a thought experiment, what would three items did you wish you had that you think might be more helpful?

Dr. Gregory Simon (08:04):

Yeah. Well, we certainly would like to know what were the circumstances that people were experiencing and how those might be related to their likelihood of improving. But I'd say I would not be completely sure that that would be very predictive. And we did, for instance, we had some very, very, you could say, crude indicators of people's circumstances. We did have indicators for instance, of the characteristics of their neighborhood. We didn't know people's personal income or household circumstances, but we did know what kind of neighborhood they came from and that was in the variables we could consider, but it didn't predict, didn't tell us much.

Dr. Lisa Dixon (08:49):

It just really makes you think about what we're missing in clinical care. If the key indicators that we find routinely in medical charts aren't helpful, I find it very sobering that it makes me feel like we may be missing important aspects of a workup or a dialogue or at least not recording them in a meaningful, predictable way.

Dr. Gregory Simon (09:16):

Well, I think having had this experience, I guess what I would say is I take away from it, and I don't think this is news, that depression is an extremely heterogeneous condition. And to be honest with you, it's probably more accurate to say depression is an extremely heterogeneous cluster of conditions which look the same on the surface. That depression is probably not a thing or as I've sometimes said, "The diagnosis of depression is a sometimes useful fiction." By that I mean it's always a fiction and it's sometimes useful. So that the thing we call depression is a lot of different things, different things in different people. And we know that depression treatments affect different people in very different ways and that we, at this point, have a pretty poor understanding about that variation and what causes it and certainly how we can understand or predict it.

(10:07):

So I would say interesting we came to this work, because we'd been doing work sort of using health records data to do prediction about other things. We'd done a lot of work about say, predicting the likelihood of self-harm events or suicide attempts or predicting opioid overdoses or predicting who would have a psychiatric hospitalization. Those things we were actually surprised at how predictable they were from the information and health records. And there's really a pretty stark contrast between predicting those really negative events that people experience from information and records and predicting people's likelihood of improving when they start depression treatment from the information in people's records. The contrast is pretty stark.

Dr. Lisa Dixon (10:50):

That's fascinating. Fascinating.

Dr. Gregory Simon (10:53):

Yeah, so when you think about, say especially for instance suicide attempts or self-harm events, there are a lot of different indicators of accuracy of prediction, but if you talk about something called a Cstatistic or area under the curve, which is a sort of rough indicator, it goes from zero to one, but 50% is flipping a coin. So really you could say it goes 50% is flipping a coin and one is perfect prediction. Many of these models predicting suicide risk are now getting up toward 0.9. While as we published these models predicting outcome when people start treatment for depression, in this case antidepressant medication, but in the previous paper psychotherapy they're more like 0.6, so hardly better than flipping a coin.

Dr. Josh Berezin (11:35):

I mean, the other side of Lisa's point also is not just that you have this heterogeneous disorder that we're calling depression, but also the uses and the design of electronic health records. What is actually captured there? I always think of them as essentially becoming a billing tool rather than a clinical tool. So in some senses it might not be that surprising that the information captured there isn't going to predict all clinical outcomes. What do you think [inaudible 00:12:09]?

Dr. Gregory Simon (12:09):

Oh, you're right. The information people's records is at a relatively crude level, certainly true, but it still is pretty notable that there's some events that can be very accurately predicted from people's health system records, even using just that crude simple information. We came into this intentionally using pretty much exactly the same sort of types of data, exactly the same predictors, even exactly the same statistical methods, coming into it saying that worked really well for predicting these other things, let's hope it works as well here. And it definitely did not. So that contrast to us was pretty notable.

Dr. Josh Berezin (12:44):

How about on the flip side for the results that you found for suicide, for example. Why would that be a more predictable event based on electronic health record data?

Dr. Gregory Simon (12:57):

Yeah. Well, I'll answer that first in terms of you could say the math of it. In both of these areas we could say if we look at say, risk of self-harm or suicide, there are a lot of things that individually are sort of moderately related. No one of them is an extremely strong predictor, but many of them are moderately related. And it turns out if you know several of those things like 50 or 80 or 100 of those things and you add them all up in the right way, they all add up to we can really predict this.

(13:32):

What we found was if we look at prediction of outcomes of depression treatment with either medication or psychotherapy, you can still see there are a lot of these individual things that are moderately related. It's not that having a history of psychiatric hospitalization is unrelated to getting better. It is moderately related. But at least what the arithmetic or the mathematics or the statistics tell us is that in the case of predicting outcomes of depression treatment, when we add all these things up together, they don't hang together and they don't give us a very strong prediction all of them considered.

(14:07):

The situation when we're using these kind of machine learning models, these statistical models in terms of predicting things in medicine in general, but in mental health is, we can point to situations where we can look at a sort of a large number of moderately correlated indicators and we can combine them in a way that's very powerful. But sometimes there are other cases where all of those moderately correlated indicators don't combine very well and the sum is really worth less even. The whole is worth less than the sum of the parts.

Dr. Lisa Dixon (14:41):

Would you characterize that as the machine isn't learning very well?

Dr. Gregory Simon (14:48):

Well, I have, as you may know, I mean I have quarrels with these terms, machine learning and artificial intelligence, because machines don't learn and intelligence is not artificial. But that's a whole other conversation that we could have. But I think the math is telling us the truth. Math does reveal the truth and what it says is, here's a situation where for instance, use that say risk of psychiatric hospitalization or risk of suicide attempt, that there probably is a coherent vulnerability, a coherent liability. There is a true thing that lies underneath it and it is maybe not one thing, but mostly one thing.

(15:26):

So that there's a relatively homogeneous thing which we hope to predict, but cannot observe, and we can see these indicators on the surface and these indicators on the surface are all pointing kind of to the same thing. So that works. In another situation we might say, "Well, the problem is, that thing is not a thing." There's several different things. There's not one thing underneath there. There are lots of different things and we may have all these different indicators that are pointing at different things. And when we try to add them all up together, they don't find anything, because there's not one thing to find.

Dr. Josh Berezin (15:59):

So stepping back for a second, how much of the methods in this paper are something that you could have done five years ago or maybe did do five years ago and how much is part of the new kind of expansion with AI and machine learning?

Dr. Gregory Simon (16:15):

Yeah, well this gets into, in this paper we use you could say, old methods and middle-aged methods and not brand new methods. We use methods that have been around for 20 years, methods that have been around for 10 or 15 years, not the ones that have been around for just five years or so, although we have in other settings say, looking at prediction of self-harm or suicide risk, we've done published work using sort of comparing all of these different generations of methods. With the data we have available and in our hands at least, the newer and I would say more opaque methods do not actually perform better. I have opinions about this and if we look at least in the mental health area, it's pretty hard to point to a situation where the newer and more opaque methods actually do perform better. There probably are situations where they do.

(17:08):

Those newer and more opaque methods might perform better, say in the machine reading of images. Where for instance, a machine might discover particular things about a picture of the back of someone's eyeball or about some other kind of image that a human being does not even know and could not find. When we're talking about the kind of, to be honest with you, more plain or apparent things we're talking about, if you look at the list of things that in our work have predicted self-harm events or suicide attempts, they're the things any reasonably informed clinician or even any reasonably informed lay person would know about. So at least in the cases that we're looking at, there does not appear to be any deep magic or deep mystery to be discovered. So the relatively simple methods work just about as well as the more complex ones. There are those who are advocates of the more complex methods. To be honest with you, there's sometimes people who have a business model of selling the more complex methods, so they would not surprisingly be advocates.

Dr. Lisa Dixon (18:10):

No, that couldn't happen.

Dr. Gregory Simon (18:14):

But I feel pretty strongly that we should use the simplest and most transparent methods possible, because when we think about use of these tools in clinical care and use of these tools to say things to clinicians or even more important to say things to patients, people will often ask the question, Can I believe that? How do I know that? Can I trust that?" And things which are more transparent are typically more trustworthy. This is of course an empirical question. This is a scientific question, does this much more complex and less understandable method perform better? That's a question to be answered by data. It's not a question of opinion. But if we ask the question and we find the simpler methods work about as well, then I think the simpler methods are definitely preferred.

Dr. Josh Berezin (18:59):

In terms of this particular question about depression outcomes, is your next step in the research to try and extract better information from the existing EMRs or make the EMRs better, or is it to do... It sounds like from everything you've said that until we have a better sense of exactly what we're talking about with the depressions or this, as you've been saying, heterogeneous group of things that we're calling depression, is the next step actually doing a better job as we've been trying to do for a very long time now, clarifying those? It strikes me that until you solve that problem, you could have the fanciest AI or machine learning and you're still going to get something that's not going to jive, because we don't really know what we're asking it for.

Dr. Gregory Simon (19:54):

Yeah. In terms of what we're doing, some of the things we're working on in the very near term are sort of relatively simple and to be honest with you, pretty apparent next steps. One of those is trying to get better information about people's life circumstances and the health systems interestingly have started to sort of systematically survey people about social determinants of health. That information is coming into people's health records, so that's information we will certainly be using. We're also very interested in stitching together how people respond to different treatments across time and both because we think that might be predictive, but also because it might help us to understand different subtypes of depression. The challenge is that even if we say this is a person who took a particular medication, medication A, and seemed to get better, a lot of the people who got better, that's not a specific response to medication A. Those are people who may have just got better spontaneously or got better with any medication.

(20:49):

But if we're able to stitch together how people responded to multiple episodes of treatment across time and say, "Well, here are people who did not respond well to medicine A and did not respond well to medicine B and responded well to medicine C or did not respond well to medicines A and B, but responded well when they received psychotherapy," those may be sort of different subtypes of depression and that might be much more useful in terms of trying either to understand how people respond to treatment or to predict. One of the other things that I'm hoping happens and will be a very interesting development, but probably a bit farther off, actually National Institute of Mental Health has recently started a large program and I think we'll be about to start up several large research projects trying to use various more sophisticated tools for parsing depression and predicting outcomes of depression treatment.

(21:38):

There's a long history of this of course, and to be honest with you, a lot of it is pretty disappointing, but what they're focused on, which I think is actually potentially quite promising, are some of these sort of, you could say, brief neuropsychological assessments or what we might call task-based assessments, both because those things actually could be used in practice. It's very unlikely that at least in my career, that I will have a very high-powered MRI machine in the corner of my office that I will ask people to stick their heads into, but that I would ask people or that our health system would ask people before their first visit, "We'd like you to do this online assessment and complete these sort of five little video game-like tasks," and that might help us to decide what the best treatment for you is. That could actually happen. So there will be a big research activity focused on that, which likely would not yield real results or insights for 3, 4, 5 years, but I consider that to be very promising.

Dr. Lisa Dixon (22:32):

Greg, I wanted to follow up on a slightly different thread, which is are there danger... Let's say you have a great predictive model. How do we actually use it in clinical care? In my own work I've seen and heard people express a lot of concerns, particularly people with lived experience, people who are receiving treatment, as to how the prediction would be communicated, how would it affect the utilization of treatment, the notion of hope. And hearing people talk about this, it did sensitize me to the fact that there might be some dangers of misusing predictive models and that we have to think about optimizing their use assuming we do have good predictive models.

Dr. Gregory Simon (23:21):

Yeah. Well, let's take this particular case. We've done a lot of work on prediction, because when you think about how these things should be used and what the potential negative consequences are, it depends absolutely on the very specific use case you're talking about. So I know there are people who say, "Oh, prediction models are evil," which is about like saying, "Math is bad." It's just about as true and just about as useful in terms of guiding practice or policy. What we want to say is what are you wanting to do with it? What is the problem you're trying to solve or the question you're trying to answer? There is, I think sometimes in our field, a little bit too much fascination with we have some data, we have an R package, let's put them together and see if there's anything publishable. That's not the right way to do it.

(24:11):

The right way to do it is to say, "Who are the people we are trying to help? What is the problem they have or the question they need an answer to? Can we help them? Can we come up with a useful answer?" So when you think about this very practical situation, typically what you're talking about is a person who's living with depression may be visiting a clinician for the first time or even deciding about what kind of clinician they would want to visit, and what they're saying is they're asking questions like, "Is it even worth the bother? And if I have different options, which one is going to be better for me?" Those are the questions we would like to answer.

Dr. Lisa Dixon (24:49):

See, I think that's so important the way you presented that, particularly if you're choosing between a couple of options. And what do the predictive models tell us about the advantages, disadvantages of A versus B, which to me it seems that's a very empowering, recovery oriented way to think about it.

Dr. Gregory Simon (25:11):

Yeah, so for instance, if we could say, let's take the two extremes. One extreme is, what if we could say to people, "You're likely to feel a lot better in about a month, even if you never took a medicine for depression or didn't see a therapist, this is going to get better." Now, to be honest with you, there might be some in our field who would have problems with that prediction, but if we could do that, that would be incredibly useful. Take the other extreme, what if we were to be able to say to someone, "It looks like you are somebody who could probably try three different antidepressant medicines and none of them would work. You might want to just go directly to some treatment that is more likely to be effective for people for whom medicines don't work."

(25:55):

Now, you might see that as a pessimistic prediction and it might be a pessimistic prediction in terms of your likelihood of doing better in a few months is less than the average person. But it's still a useful prediction if we say, "Don't waste your time on this, don't waste your time on that. You might want to go straight to the thing that we tend to do after three things haven't worked, because it looks like those three things won't work for you." We are not there and we're a long way from being there, but I think even a so-called pessimistic prediction can be helpful if it considers alternatives and says, "Maybe none of these alternatives are great, but this one looks like the best one for you."

Dr. Josh Berezin (26:33):

If you had a successful predictive model, I'm assuming it would come up with a percentage. People with your characteristics have about a 30% chance of improving on medications, is that right? Is that what would come out of that?

Dr. Gregory Simon (26:51):

Yeah.

Dr. Josh Berezin (26:52):

In that case, does it sort of give people like, oh, well that's not zero, it's not 90, but I now kind of know what the odds are and whether or not this might be worth pursuing first or second or third?

Dr. Gregory Simon (27:06):

Yeah. And there are very important health communication questions, but also some established knowledge about health communications, about how we communicate effectively to people about things like probability. I think as I sometimes say, people can understand probabilities if you talk to them about it in the right way. If you watch the evening news, there's this whole section on sports and there's this whole section on weather, both of which are all about probability. And people can understand what a 40% chance of rain is, a 90% chance of rain, because they continuously receive this information and they correlate it with their experience and they learn something about that. Not everyone is sports fans, but to be honest with you, more sports fans, they can understand probabilities really well. So people can handle this. It's not that this is impossible. The health field has not always done a very good job of this.

Dr. Lisa Dixon (28:05):

I do think that it's important to be proactive and intentional, I think, in identifying the challenge and trying to meet it. It's a big responsibility.

Dr. Gregory Simon (28:20):

One thing I think that to me is a very interesting question when you think about... It's interesting because we're starting this discussion in a case where we're saying, "Boy, fancy prediction didn't work here." But we're presuming that well, this is going to get better and we are going to be able to predict things better in the future, and what does that mean and what does that mean for healthcare? What does that mean for people who live with mental health conditions?

(28:44):

I'm a bit of a weather nerd since I'm a skiing fanatic, so I pay a lot of attention to the weather. But in the weather world actually, and I think most people are even aware of this, there are two very different disciplines and professions. There are the atmospheric scientists and there are weather presenters. Atmospheric scientists they are serious nerds. There is nobody who's deep into hairier math and complex data spaces than weather modelers. But those are probably not the best people to talk to you in the evening news about what this means for you and how you're going to make decisions in your life based on it.

(29:24):

There are weather communicators, most of those people, to be honest with you, did not major in math. In the weather world these roles are fairly distinct. I think we've got some interesting future in the healthcare world about whether those roles remain within the same person or how they differentiate. I don't think it's the wrong thing for clinicians to think of themselves as the weather presenters rather than the data crunchers. That is a very important and useful function, which likely will never be replaced by a machine. To be able to sit with someone and say, "Well, all of the information we have available tells us that these are the kind of things that would be more likely to occur, and let's talk about what that means for you." Ideally, we'd be able to tell people in a fairly sophisticated way about these are the good things that might happen and here's how likely they are and these are the bad things that might happen, and here's how likely they are.

(30:24):

Being a bit of a nerd myself, as I sometimes say to my patients when I'm trying to convey to them sort of what the evidence says about the clinical decision we face, what I'll sometimes say is, "It is my job to be able to know about the probabilities. It's your job to know about what matters to you. That's not for me to decide, and that's certainly not for me to know. That's for you to know and to tell me what matters to you." The idea that I might be able to say the best information we have says if we were to do this, I'll try to put numbers on it, but in a way that people can understand.

Dr. Lisa Dixon (30:56):

That's just so well said. Thank you. I personally find that to be very helpful.

Dr. Josh Berezin (31:02):

Yes, and sometimes you get the rare person who is both the weather nerd and the good presenter, and it seems like you're kind of falling in that really unique space.

Dr. Gregory Simon (31:14):

Well, I try to move back and forth between those two worlds. They're interesting worlds, but they involve somewhat different... They are different. Those are different skills, I think.

Dr. Josh Berezin (31:22):

Well, that seems like a good place to stop, and we just wanted to thank you so much for the paper and for coming on the podcast today with us. It was a truly fascinating conversation.

Dr. Gregory Simon (31:32):

You're most welcome, enjoyed talking with you.

Dr. Lisa Dixon (31:35):

That's it for today. Thanks to Aaron van Dorn for mixing and editing and Demry Jackson for additional production support. We invite you to visit our website ps.psychiatryonline.org to read the article we discussed in this episode, as well as other great research. We also welcome your feedback. Please email us at psjournal@psych.org. I'm Lisa Dixon.

Dr. Josh Berezin (31:57):

I'm Josh Berezin.

Dr. Lisa Dixon (31:58):

Thank you for listening. We'll talk to you next time.

Speaker 4 (32:02):

The views and opinions expressed in this podcast are those of the individual speakers only and do not necessarily represent those of the American Psychiatric Association. The content of this podcast is provided for general information purposes only and does not offer medical or any other type of professional advice. If you're having a medical emergency, please contact your local emergency response number.