Opening the "Black Box": A Conversation with Microsoft's Rich Caruana About AI in Health Care

Interview conducted by Sean Sylvia

Introduction

Rich Caruana, senior principal researcher at Microsoft Research, has been working in machine learning in health care since the 1990s, when his graduate advisor asked him to help train a model to help a group of clinicians and scientists predict pneumonia risk.

Caruana knew that even if he could create an accurate model for predicting pneumonia risk, it still might not be safe and effective to use with patients. Other similar models had produced erroneous results, such as asthma presenting as a *protective* measure against bad outcomes of pneumonia. Previous research had explained this as a real pattern in the data with a nuanced explanation that the artificial intelligence technology couldn't decipher: people with asthma were more likely to seek care for pneumonia as soon as they noticed symptoms, and to receive high-quality care.

"This has always bothered me, as I continue to work in machine learning for health care, that the most accurate models tend to be 'black box,' uninterpretable models," Caruana told **Sean Sylvia, PhD**, of UNC for the *North Carolina Medical Journal*. "Statisticians have now figured out how to build a class of machine learning models that are just as accurate but are completely interpretable and understandable to doctors. It puts the human expert back in the loop."

Now that machine learning models can be created using a "glass box" framework that allows users to more deeply interrogate the data, clinicians and policymakers can better understand not just that they work, but how and why, allowing more effective and tailored applications.

This interview has been condensed and edited for clarity.

Sean Sylvia (NCMJ): Why does it matter to be able to "see inside" a large language model or machine-learning algorithm that is being used in a health care capacity?

Richard Caruana (Microsoft): *The most important thing, I think, is that every model ever trained with machine learning on any medical data set has learned a half a dozen things which are wrong. And by "wrong," I don't mean statistically untrue—they are statistically true in the data, but they're wrong given what you want to use the model for. If the model is going to decide whether you should be admitted to the hospital, or whether you should get a certain treatment or not, it is just doing statistical association. It would predict that a 105-year-old with a history of asthma, heart disease, and now presenting with a blocked airway is actually low risk for dying of pneumonia, because those patients usually get very high-quality treatment. In another example: if you have a policy that every patient admitted with certain kinds of heart disease gets a baby aspirin each day, and you follow that policy religiously, you'll never see a patient who didn't receive the baby aspirin and will never be able to estimate what the risk would have been if that happened. You can run into a problem that you'll never see a statistical sample of patients who don't receive the same protocol, and the model will not be capable of learning something that is causally correct because it is only observing one arm of a protocol—when it goes right.*

"An AI model could be providing a safety net, double checking things and raising a flag whenever something looks questionable and explaining why."

These effects I'm describing have been in all models, but no one knew they were there. Now that we have "glass box" models where you can see these things, people are asking if the policy should be that you have to use a model like this in health care any time you are working with tabular data. The community is waking up to this, and it's very, very important.

Electronically published July 10, 2024.

Address correspondence to

N C Med J. 2024;85(4):254-255. 0029-2559/2024/85410

NCMJ: I hear you saying that there's increasing realization that AI predictions aren't necessarily the values you need to guide you in a medical context. How do you think about generating actionable insights from machine learning models in health care?

Caruana: *What we've found is that every rich complex data set has a lot of cool things in it, and if you can train a high-accuracy, high-interoperability model on it, you can learn a lot. Sometimes we just want to learn about the data and don't have any intention of using the predictions to do anything. I might give you 100,000 patient records and you can't wrap your head around that much data, but training a model on that data might be useful. When you use an interpretable model, you have something that a data scientist and a clinician working together can dive deep into and get a very nuanced idea of what's happening between different groups and their care outcomes.*

I don't want to anthropomorphize, but we are getting to the point where you could imagine a physician communicating with these things and benefiting from them almost as if there was a second physician providing a second opinion looking over their shoulder at every major decision they're making. An AI model could be providing a safety net, double checking things and raising a flag whenever something looks questionable and explaining why. I believe physicians will be very happy with that if it's done well, if it doesn't alarm too often.

NCMJ: How do you involve clinicians and policymakers in the development of these kind of learning models?

Caruana: *It's critical that clinicians are involved early in the process of training a model, helping to answer questions about how to collect the data and where to collect it from. We often do a discovery phase, and then a vetting phase where we deploy the model clinically.*

We clean up the model and then we deploy it, and the truth is that many end users don't want to ever look at the model. They just want to know the standard traffic light, "red, yellow, green" kind of thing. Only if the model really seems to strongly disagree with a decision they were about to make would they even want to click on something. It's working in the background. Once you deploy a model, you want to have documentation just like you would have with a new medical procedure or drug, about how often the model is false alarming, missing things, or actually making a difference.

"It's critical that clinicians are involved early in the process of training a model, helping to answer questions about how to collect the data and where to collect it from."

NCMJ: What advice do you have for interpreting the findings of models like this and adapting to their use in the future?

Caruana: *There are surprises in all data sets, and it is critical that you can understand what your model is zeroing in on, because some of that will be exactly right, and sometimes it will be zeroing in on something that is a treatment effect or something else you don't want it to learn to serve your purpose of learning the risk before treatment, for example.*

Now that we have interpretability, there is no going back. If you're building the Brooklyn Bridge, and all the engineers and builders are blind and you suddenly give them glasses and they can see, nobody is going to take the glasses off willingly and go back to building the bridge blind. I know clinicians now who say they will never use a non-interpretable model again because we have a better option now. NCMJ