Ziad Obermeyer, MD:
Some people are doing amazing work and there's so much creativity. I think there's so much opportunity to bring data to bear on some really important questions in medicine. On the other hand, the fact that there was a hugely biased algorithm that was not only made, but scaled up 70 million people a year and nobody caught it, suggests that there is a market failure here and that there is a clear role for regulation.

Ellen Kelsay:
That's Dr. Ziad Obermeyer, Blue Cross of California Distinguished Associate Professor of Health Policy and Management in the School of Public Health at UC Berkeley, where he conducts research at the intersection of machine learning, medicine, and health policy. Ziad was named an emerging leader in health and medicine by the National Academy of Medicine and has received numerous awards, including the Early Independence Award, the National Institutes of Health's most prestigious award for exceptional junior scientists. In addition to teaching and conducting research, Ziad is a practicing emergency medicine physician.

I'm Ellen Kelsey, and this is a Business Group on Health podcast, conversations with experts on the most relevant issues facing employers. Today I sit down with Ziad to discuss artificial intelligence and health care, including its perils as well as its potential.

Ziad, welcome. Thanks so much for joining us today.

Ziad Obermeyer, MD:
This is great. I'm really looking forward to the conversation.

Ellen Kelsay:
We are as well and we're just so interested in your research and work. Let's really dive right in. We certainly do hear a lot about artificial intelligence and machine learning, especially in health care these days. For those of us who perhaps might not be as familiar with what that is or fully understand AI and how it works within the health care field, can you help us level set a bit?

Ziad Obermeyer, MD:
Sure, I think it's a term that covers a multitude of sins and I think the broadest possible definition, which might not be the most useful, is that it's a set of tools that helps you find patterns in data. I know that seems overly vague, so I'll just give you an example of one of the things that people commonly use this for, which is in health care we are always happier if we can prevent all this from happening rather than treating it after it happens. One of the things, for example, that machine learning has proven very useful for is saying, okay, I'm seeing a patient in my clinic or in the emergency room and I want to know something about how that patient is going to do in the future. Is that patient going to spiral downwards? Are they going to do okay? If they're spiraling downwards, there's some stuff I would do for that patient now. Machine learning can be used for those kinds of problems where you're trying to look ahead in the future and find signals in the patient's data that you already have when you're in the clinic or in the ER, and use those data to make a prediction about what's going to happen to that patient so that you can make a decision better than you otherwise would be able to do. It all rests on that idea of finding these hidden signals in the masses of health care data that we have available to us and using that to make a prediction.

Ellen Kelsay:
That's great. That was a very simplified, understandable answer. I appreciate that, because behind the scenes it's very technical, but from a cursory understanding, maybe not as technical as some of us might think it is. One thing I know that some hear, and perhaps it's a misperception, is that AI replaces the clinician. I think in what you just said it's not necessarily replacing the work of the doctor or the clinician, but it's really helping to support them in diagnosis of disease. Would you say anything more related to

that and any concerns that people might have about a machine doing the care of a doctor versus it supporting the care of a doctor or clinician?

Ziad Obermeyer, MD:
Yes, I certainly think of it the way you do as an adjunct to the doctor. I think a lot of the reason that I do that is very germane to our conversation that I'm looking forward to later on about bias. The reason those two are so related is because we almost never have in our data the exact thing that we want. When I tell you, oh, is this patient going to spiral downwards in terms of health, there's no variable in my data set called spiraling downwards in terms of health. There's just a ton of other stuff that's valuable, but that actually really needs a human to make sense of and to catch any problems that might arise from the fact that we don't have the exact thing we want. In many ways, I think this will be familiar to probably many of your listeners, because it's often the case in life that we want to measure something. We want to measure quality or we want to measure an employee performing well, but what does that mean? What's the variable in my data set that captures whether an employee is performing well or whether this policy is targeting the right people? All of those problems with measuring the right thing, actually make it really hard for algorithms to predict the right things. Algorithms can be incredibly useful in the hands of a human who can put them in context and get out of the algorithm what she wants without falling into some of the traps that algorithms often fall into.

Ellen Kelsay:
That’s really helpful. Really it does sound like AI is here to support and augment and really to improve health care overall. I imagine there are lots of different ways when you kind of define improve health care that could be potentially improved patient experience, less waste and duplication of services, potentially even reduce costs and hopefully improved outcomes. Would you say that AI has an effect on all of those things?

Ziad Obermeyer, MD:
I certainly hope so. Despite a lot of my work, which is about where algorithms go wrong, most of what I believe is that algorithms are going to be a transformative and very positive force in the future of medicine. I'm very optimistic about that. I, of course, believe we need to avoid all of the pitfalls that come with doing anything new, because these are tools that can harm just as easily as they can help. But yes, absolutely. I think they're going to be really transformative in terms of helping doctors make better decisions. I can give you just a thumbnail sketch of one example from my work where we look at how doctors make decisions about testing people for heart attack when those people come into the ER. One thing we know about how doctors test is that we test way too much and we see that because a lot of those tests come out negative. Once you have the negative test, it's easy to say I shouldn't have done that test because it came back negative and now I've exposed the patient to the financial costs and the health risk of the test and I'm no further in my knowledge than I was before. What we also found is that doctors, in addition to over testing low-risk patients that an algorithm could have told you not to test, doctors also tend to fail to test some high-risk patients and those patients go on to have very bad outcomes over the next few days after they are, for example, sent home from the ER. The reason I like that example is because you can see how we don't need to test more, we just need to take some of those tests that doctors are currently doing on low-risk patients and reallocate them to high-risk patients. So our testing budget is the same or it can even be lower, but the value of testing is much, much higher. I think that's a general example that cuts across many, many different parts of health care. Doctors are human, they make mistakes and algorithms can help them avoid both kinds of mistakes, overuse in low-value people and under use and high-value people.

Ellen Kelsay:
That was a great example and thank you for sharing that. You mentioned that your life's work is really looking at sometimes when these things can go wrong. Let's talk about potentially the pitfalls or the dark side of using artificial intelligence in health care.

Ziad Obermeyer, MD:
Just to be clear, there are a lot and it's why I work on that topic so much. I think that the thing that we have to remember when we're training algorithms is that algorithms are, as I mentioned, they are ways to find patterns in data, but all of our data that we have available to us to train those algorithms are filtered through a health care system that is itself biased and unfair and that health care system is the product of a society that is itself also biased and unfair. The very data on which the algorithms are often learning have these huge landmines lurking in different parts of the data where if you're not careful, those algorithms can scale up and reproduce all of those ugly things that we don't like about our health care system, because they're in the data. The algorithm is just doing what we've asked it to do, but it's learning from a biased data set in a biased society and health care system. I think that's the big picture for why we should really care about this. It's that not only is the algorithm learning our own biases and errors, the algorithm is as a result not doing what it's supposed to do. We often want algorithms to help us target tests and other resources to people who need them most and because the data aren't an objective readout of who needs health care most, they are a readout of who gets health care. There's a huge bias built in.

Ellen Kelsay:
How do we fix that? How do we fix the underlying flaws in the data set and the underlying flaws and the people amassing the data so that the algorithms do work and do leverage the most effective and appropriate data sets?

Ziad Obermeyer, MD:
It's exactly the question that I think we all want to answer. Should I maybe give you a little example of one algorithm that we've studied? I think it highlights really all of these questions that you're asking about, how we get into these messages and how we might get out of them. Anyone in health care will have heard about is the crisis of affordability, which is largely driven by a small number of very high cost and high complexity patients. For the past decade or so health systems have been working intensely on that problem, trying to find those high-risk, high-need, high-cost patients and get them the help that they need with their health. Everyone wins and the patient avoids exacerbations of heart failure or strokes or things like that and the health care system wins because those patients no longer need to go to the ER and end up in the hospital and rack up these huge bills. This high-risk care management or population health management or care coordination, as it's variously known, is really a cornerstone of how we're dealing with the affordability and the spiraling costs of health care in this country. Those programs are not cheap. The goal is to improve health and to save money, but sometimes to save money, you have to spend money. Those programs involve a team of highly trained nurses, usually who can make house calls, who can spend time on the phone whenever you need help, who can find you a primary care appointment on short notice. All of those things are expensive, so you can't do this high-risk care management for everyone. You can only do it for the high-risk people and that's where the algorithms come in.

If you take the estimates that the kind of analytics industry produces in health care, algorithms are being used to make decisions about population health management for 150 million people every year in the U.S. That’s the majority of the population being screened in one way or another by these algorithms. We studied one of those algorithms that was made by one company, but it's a very general thing. Almost every algorithm in the space looks the same. They're doing the same thing and they're doing something related to what we talked about earlier, which is to say, okay, I've got a primary care population and I know that there are some needles in this haystack that I need to find today because they're going to get sick tomorrow if I don't do anything. It's a very noble goal. We want health systems to be proactive and to be looking for patients who are teetering on the edge and intervening early. This is all very good and this is exactly the kind of job that we would want an algorithm to do because primary care doctors are busy. They don't have time or even potentially the expertise to look ahead and see who's going to be spiraling out of control. This is a great use of an algorithm, but how do these algorithms work? Well, they say, okay, we're sitting here today and I'm going to look ahead over the next year for all of your primary care patients. In the setting we studied, this was not a huge practice, but it was 40,000 people in a large academic hospital's primary care practice, so the algorithm sifts through all of their data on these tens of thousands of people and makes a prediction about who's going to generate a lot of health care costs over that next year, as a measure of people getting sick, because of course when people get sick, they end up in the ER or they end up in the hospital, they generate all these health care costs.

When we studied this algorithm, we looked at black versus white patients in this primary care population and we looked at two patients at the same score. You can imagine, when we take a sample of the population, people of the same algorithm score, the high-risk people get automatically enrolled into the program, basically the top half gets shown to their doctor and the doctor says yes, this person would benefit or this person wouldn't, but the algorithm score is kind of dictating what's going to happen to that person if they get screened in, screened out, or if their PCP decides. When we looked at two patients at the same score, one black, one white, those patients should have the same health care needs because they're being treated the same way, but they did not. When we looked at the year after the algorithm made its prediction, we found that the black patients went on to have far greater health care needs than the white patients, even though the algorithm was treating them the same. Another way to think about that is imagine all these primary care patients lined up and in order of their algorithm score and the front of the line gets to be admitted into these extra health programs. What that algorithm was doing is it was letting healthier white patients cut in line ahead of sicker of black patients.

When we looked at the racial composition of that high-risk group that got automatically enrolled, it was far less black than it should have been had we been purely concerned with health care needs. For us this was a pretty clear example of bias that was induced by an algorithm. Why did that bias arise? Well, actually, I already gave you the clue, which was that the algorithm was predicting not someone's health care needs, but their health care costs. Even though those two things are very related and it's very tempting when we're talking about health care to reduce needs into how much care does someone receive, how many dollars did that cost, that was the source of bias in this case. Because of barriers to access, because of structural inequality in our society, even in this pretty well insured primary care population at an academic hospital, black patients generated fewer costs at the same level of health.

If you think about two patients with the same set of illnesses, like heart failure and hypertension and diabetes, we can kind of equalize all those things, on average black patients with those illnesses cost less than white patients. Why, because when you need health care, you don't always get it and some people are less likely to get it than others. Some people, even when they make it into the health care system, they get treated differently than others. There are a number of studies on this point, but I think one of the more elegant ones is one that was in the *New England Journal of Medicine* a few years ago. They had little vignettes of patients who might need some testing for heart attack. They standardized the vignettes so that the vignettes were exactly the same, but they randomly changed the picture that accompanied the vignette to either a black patient or a white patient and cardiologists were 25% less likely to recommend catheterization for black patients, even though the other information was exactly the same.

There's all sorts of evidence that we treat patients differently based on the color of their skin, based on their education as well, based on their socioeconomic status, but especially based on the color of their skin. All of those things mean that black patients cost less than white patients, even when they shouldn't, even when they need as much health care. That's what the algorithm was seeing. We asked the algorithm to predict the costs as a proxy measure for needs, and that's exactly what it did. We didn't realize that that request was the source of the bias. It encapsulates a lot of the things that I've seen in other places in my work where we blame algorithms for the bias that they perpetuate and scale up, but in fact, the algorithm is just a mirror. It's just us. We asked the algorithm to do the wrong thing and that often just results in the kinds of biases we see.

Ellen Kelsay:
That's a really big problem, right? You said it's perpetuating a bias and it's perpetuating a big problem that we know exists and we're all talking about equity and health equity and achieving equity for all, we need to find a way to not have the algorithm perpetuate that issue. I'm curious with this anticipatory algorithm that's not based on cost, but really is based on need, are there ways to tell the algorithm to calculate an anticipatory event that hasn't yet occurred?

Ziad Obermeyer, MD:
Yes, in fact that's exactly what we did. When we first saw these results, we were surprised and we actually reached out to the company that makes this algorithm and we said, hey, you don't know us, but we've identified this problem in your algorithm. They were incredibly responsive and very motivated to work with us on ways to fix it. What we did is we worked to put together a set of measures that capture a patient's health rather than their costs. I think it's hard to in retrospect now that you know the ending of the story, it's easy to say, oh well, like of course cost is biased and why were they using cost? The fact that many companies made the same mistake, in fact, there are academic research groups and parts of the government that use a similar strategy for doing what they call risk stratification. Hospitals, many of them non-profit, many of them with deep commitments to health equity, bought and applied the algorithm. Doctors didn't overrule the algorithm. This was a subtle seeming problem that just had huge consequences for patients, but it wasn't caught. It wasn't caught because cost is actually very correlated with needs. It's just differently correlated for black and white patients. It's also a very appealing variable to use for all of us working in health care, because it's in claims data, when it's missing it's zero. It just has all these appealing features that make it very tempting to use as your summary statistic for health. As we know, health is not simple. There's no summary statistic for health and you can only get it, the concept that when we're talking human to human, we can say health and you know what I mean, but when you're trying to train an algorithm, you have to triangulate health across a number of different variables. If you want to use laboratory studies, you have to deal with the fact that not everyone has a laboratory study. So what do you do with those people? It's just a lot more complicated to try to train an algorithm to predict health, but I think the return is huge. Finding these less biased measures on which to train algorithms is a very, very high-value activity.

That's what we did. We worked with a company and in their own data, we retrained the algorithm to predict things that were more aligned with health than with costs. That dramatically reduced the bias. We got a lot of publicity after that study, which was great. We tried to turn that into more relationships with health systems, with insurers, with parts of the state and federal government with regulators number one, so we could help them as they tried to diagnose and fix the bias in the algorithms that they were already using, but also so that we could learn kind of the taxonomy of biases that affects all sorts of algorithms in health care. We're kind of doing a bunch of that work now and we're going to distill those lessons into what we call a playbook, which is the kind of document that if you're the CEO or the COO of a company and you're worried about bias, this is a document that you would be able to just send to your chief data officer or to your CTO and say, do this. Then with some allocation of time from a technical team, you could work through an inventory, all the algorithms that are in use, grade them all for bias and then put in place the steps to fix them.

Ellen Kelsay:
That's great. I bet this playbook will be immensely valuable to just about anybody who is a party to the health care delivery system. You mentioned a number of parties just now. You mentioned this one vendor that you were working with in the specific example to help them retrain their algorithm. You mentioned the non-profit hospital. You mentioned government and regulators. If you could maybe pull forward one or two recommendations from that playbook that you could share or are currently sharing with those that you're working with that are really the couple of nuggets that are most important as they think about how they are applying and utilizing AI within care delivery or their practice or their systems, what would those two things be that you would offer up?

Ziad Obermeyer, MD:
Yeah, absolutely. I'll just tell you the first two steps of the playbook and I think those are maybe the two most important ones. The first is to do an inventory of the algorithms that are actually operating. I think one thing that was really surprising to us is how, because, of course, all businesses are complex and businesses in health are particularly complex, but each division of a company is going to be using and applying algorithms that other divisions and the upper management have no idea are operating. There's just this huge value to creating that inventory and putting it in one place and just getting everyone on the same page around what are we actually using algorithms for in this business? That's already just a very powerful thing to do because it gives everyone the opportunity to take a fresh look at what we're currently doing when normally that kind of knowledge is pushed pretty far down to a technical level at a company, because they're technical products and they're built by technical teams. Just because a product is built by the technical team, doesn't mean it shouldn't have some strategic oversight. This isn't a particular point about bias. It's just that algorithms are so central to business strategy everywhere that where an organization is using algorithms should just be a thing that everyone knows in management that needs to know about it. That inventory is really the first step.

The second step is actually, it's simple enough to say, I think it takes some practice to think about, but what we try to do is to say, okay, here's what the algorithm is supposed to be doing. In our example, it's supposed to be finding patients who are going to benefit from high-risk care management programs. Okay, great, that's the decision that it feeds into. Now, what variable literally is that algorithm predicting? The first instinct is to say, oh, it's predicting risk, but risk of what? Getting very granular is very important, because that lets you immediately see costs, it's predicting costs. But wait, costs aren't the same as a need, because some people who need health care don't get health care, so how could cost to be a good readout of someone's need? Articulating that subtle, but really critical difference between what we ask the algorithm to do and what we want the algorithm to do, is I think the foundation of a lot of the fixes that we've put in place to biased algorithms. It's articulating that difference between what we asked the algorithm to do and what it's actually doing.

Ellen Kelsay:
There's a lot that I want to ask you about related to standards. I mean, clearly there are ethical standards, there are also just kind of programming standards and consistency. I imagine, you know, is there a governing body or is each entity and organization trying to do this and create the right algorithm for the population that they're serving? I guess I would ask a couple of questions related to standards, both from an ensuring consistency and that we're not making this up as we go, and also that things are being applied from an ethical perspective. I would love to hear your thoughts on the standards question.

Ziad Obermeyer, MD:
It's a great question. I think the more I've learned about the space, I think the more, it's not surprising, it is a little bit shocking I guess. One of the reasons I like working in this space is because it's a frontier. It's the Wild West and frontiers are very exciting places to be, but they're also very dangerous places to be. I think that that's my basic impression of this field. Everyone is doing something slightly different and some people are doing amazing work and there's so much creativity. I think there's so much opportunity to bring data to bear on some really important questions in medicine. On the one hand, that's great and we would never want regulation to stifle that creativity. When I'm in my role of building algorithms that I think are good, there are lots of ways that regulation, if applied in a heavy-handed way, could make my life much more difficult and could lead to me never being able to touch the data that I need to develop an algorithm that I think is good.

On the other hand, as that example we talked through showed, the fact that there was a hugely biased algorithm that was not only made, but scaled up in the case of that particular software package, 70 million people a year and nobody caught it, suggests that there is a market failure here and that there is a clear role for regulation. I, unfortunately, can't give too many details about the particular things that I'm working on with state and federal regulators, but the thing that's been really wonderful to learn in my conversations with them is that based on their understanding of the law, unless discrimination was done intentionally by an algorithm or by the people who made the algorithm, it would be very hard for them to imagine a company, for example, being sued for having an algorithm that contained bias, if that bias was not intentional. Where companies can get in trouble is not fixing algorithms once they know or suspect that they're biased. This is a general principle of it's not the crime, it's the cover up, that gets people in trouble. I think that's very, very true here. If there's one thing that I would love to convey to a broader audience, it's that principle that you're not going to get in trouble for doing the audit and doing the investigation, but you can get in trouble for not doing that.

Ellen Kelsay:
Yes, you can't turn a blind eye to it, especially when as you were saying, there are examples of where it has gone wrong and wasn't built correctly. I appreciate that point for sure.

Ziad Obermeyer, MD:
Absolutely. Now it's pretty widely accepted that these cost-prediction algorithms contain bias for all the reasons that we talked about. Now, if you're a company using one of those algorithms and turning a blind eye to the bias, that's the problem. It wasn't a problem that we were using it before we all knew that they were biased. It will be a regulatory problem for specific companies if it's not fixed.

Ellen Kelsay:
That's a good thing. The frontier and the Wild West, like you said, is both exciting, encouraging, new dawn of exploration and discovery, but there is a lot of risk and potential unintended consequences if not done thoughtfully and appropriately. Let's end on a bright spot and let's share your favorite example of how you think AI is really being used for good and it do even more in the future.

Ziad Obermeyer, MD:
Being cautious and wanting to only speak about things that I know about means that I end up talking about my own studies, but there are just the ones that I know so I'll give you another example which is a very different setting. In this setting we're dealing with pain. Pain is something that, as we all know from the opioid epidemic, is very common in society, but it's more common in some groups than others. If you look at, for example, the fraction of people who say that they're in severe pain over the last 24 hours or whatever, that's just much higher in black patients versus white patients and poor patients and less educated patients. One obvious explanation for that is like of course those people have more things that hurt.

We, in this one paper, studied the case of knee arthritis. If you look at knee arthritis, it is more common for example in black patients, but even when you take that into account, if we look at two patients whose knees look exactly the same to our radiologist in terms of the degree of arthritis, black patients on average report far more pain. What we were interested in was given how we have formed our medical knowledge about arthritis, there's a clear potential for bias in the sense that our medical knowledge doesn't contain the experiences of diverse populations. Everything we know about arthritis and all of the grading systems that we currently use come from studies done in the 1950s of coal miners in England and comparing coal miners to office workers. This is just like a hundred percent white male population. It's very plausible that in that population, there just weren't the same things that we'd find in populations that doctors are currently seeing today, but our medical knowledge hasn't really caught up. What we did was we trained an algorithm to try to help with that. One thing to point out, and this comes back to our usual theme of proxy measures versus real measures, usually in these machine learning exercises, what we do is we train the algorithm to predict what a doctor would have said about the x-ray. In this case, of course, that's the exact opposite of what we want to do, because the whole problem is that the doctor might be biased, not intentionally biased, just biased in the sense that she doesn't know the causes of pain that affect diverse patient populations, because she didn't learn that in medical school or in residency. Instead we trained the algorithm to predict not the doctor and her opinion, but the patient and the patient's experience of pain in that knee. What we found was that algorithm did a much better job of explaining pain in everyone, but it did a particularly better job of explaining pain in black patients whose knees looked fine to the radiologist, but did not look fine to the algorithm. By training the algorithm to listen to the patient, rather than listen to the doctor, we actually used it to uncover sources of pain that were disproportionately affecting black patients and, to a somewhat lesser extent, lower income and lower education patients. I think that's really important because when you think about what we do about knee arthritis, if you go into a doctor and you have a lot of pain in your knee, the doctor is going to send you for an x-ray and if the radiologist says, well, I don't know where the pain is coming from, but the knee looks fine, then you're going to get referred to, who knows, something that is not about the knee. If your knee looks really bad and you have a lot of pain, you're going to get referred to an orthopedist and you're going to get considered for a knee replacement surgery, which can be really life-changing, but only if the problem is in your knee. What we found is that doctors are missing patients, disproportionately black, who have a problem in their knee, but it doesn't fit the mold of things that we're used to looking for on the x-ray. That study I don't think is definitive at all. We're working actually with a couple of health systems to try to form more prospective measures and do this in a clinical setting where we can get a lot more eyes on it. I think that shows how, if you train algorithms on the right variables, if you're careful about realizing which variables are biased and which ones aren't, you can really turn algorithms into a tool for breaking down those disparities, rather than for reinforcing them.

Ellen Kelsay:
Your work is just so fascinating and that study on pain was just such a great example. I appreciate you sharing this. Clearly, we are just at the beginning of this work as an industry, we still have a lot of development and discovery yet to come. It will be fascinating to watch what you and your team and others do in this space over the years ahead. As you said, you're working with regulators and others, so I'm sure there'll be even similar developments from that perspective.

Ziad, thank you so much. I really loved the conversation and appreciate you sharing all those great examples and important considerations with our audience today. So, thanks again.

Ziad Obermeyer, MD:
Thank you. It was such a pleasure.

Ellen Kelsay:
I've been speaking with Dr. Ziad Obermeyer, Blue Cross of California Distinguished Associate Professor of Health Policy and Management in the School of Public Health at UC Berkeley. For more information on this topic and other research Ziad is leading, please check out his website at http://ziadobermeyer.com/.

I'm Ellen Kelsey and this is a Business Group on Health podcast, conversations with experts on the most relevant issues facing employers today. Please consider sharing and liking this podcast.