0:00:06.1 Sarah Crespi: Welcome to the Science Podcast for August 13th, 2021. I'm Sarah Crespi. Each week, we feature the most interesting news and research published in Science and the sister journals. First up this week, investigative journalist Charles Piller discusses a vitamin D trial that may have put asthmatic children at risk for broken bones. Next, researcher Birhanu Eshete talks with producer Joel Goldberg about the dangers to proprietary data, government systems, and patient information, when the outputs from machine learning are released on the internet.

0:00:42.3 SC: Now we have Charles Piller, he's an investigative journalist for Science. He wrote a feature story this week on a common research design that can expose patients to serious risks. Hi, Charlie.

0:00:53.0 Charles Piller: Hey Sarah, nice to be here with you.

0:00:55.2 SC: Yeah, thanks for coming on to talk about your story. The main example you use to illustrate this type of research design is a clinical trial that started in 2016 on vitamin D supplements for kids with asthma. Can you describe how the Vit-D-Kids Trial worked?

0:01:13.4 CP: The trial involved a test of a hypothesis that high dose of vitamin D could reduce the number of asthma attacks kids might have. The high dose group was paired against a placebo group, a group of kids who received a sugar pill, instead of a vitamin D pill. And the issue was that all of the children were recruited because they had low levels of vitamin D, many of them frankly deficient in that vitamin, which is vital for growth and development.

0:01:48.2 SC: Right. So what argument did the researchers make for giving these vitamin-deficient kids a placebo instead of something different?

0:01:56.4 CP: The idea here is that vitamin D testing is not a routine thing in most communities and consequently, these kids wouldn't necessarily have learned that they were deficient in vitamin D, and therefore it was okay not to treat them.

0:02:11.2 SC: Yeah, that kinda goes against the trend of leaving communities better than you found them, if you're gonna do research there.

0:02:18.2 CP: Well, that's what critics certainly honed in on, they were very concerned that in essence, they were using these children and they were mostly Black and the vast majority, children of color, using them as essentially as guinea pigs in an experiment, in which the quality of care they received is not up to the standard. The standard being, if you know a child is deficient in vitamin D, you provide vitamin D supplements to that child, you do not give them a placebo under any circumstances.

0:02:50.9 SC: And what you could do in a research trial then is compare this high dose that they wanted to see if that could reduce asthma, with a smaller dose that just brings them up to recommended levels of Vitamin D.
0:03:02.9 CP: Exactly, right. And that would be the standard used in similar studies, both related to asthma and other respiratory problems.

0:03:11.7 SC: You mentioned that many of the kids recruited to the study were Black. Was there a reason for focusing on these particular children?

0:03:19.6 CP: When researchers do a study that is particularly oriented towards addressing a particular vulnerability or a disease that affects Black children, it would make perfect sense to have more Black kids in the study. The problem in this case is that even though asthma afflicts Black children more commonly, there's nothing in the study, there's no stated intent why the study addresses the issues associated with Black children in a direct way or in a way any different from any other trial.

0:03:54.9 SC: We're basically talking here about things like social determinants of health, living in polluted areas, being exposed to more particulates in the air, these might be linked to asthma and those were not part of the study.

0:04:07.1 CP: And so the problem is the majority of the risk of this experiment fell on the shoulders and on the lungs of Black kids instead of being distributed equitably throughout the population. Because there was no special kind of aspect of the experiment oriented towards solving the problems of Black children, critics were wondering, what's the justification for providing most of the risk to Black children?

0:04:33.7 SC: Despite what many call 'ethical lapses' in the trial design, this was approved for Federal funding, this was approved by IRBs, ethical review boards and then finally, the results were published in a prestigious journal. How did all this happen?

0:04:50.5 CP: Yeah, it's a bit of a mystery. To be honest with you, some of the lapses that were assessed by a variety of medical ethicists, and also a whistleblower who initially looked at this case, seem egregious by their standards and seem egregious to the people who have studied this sort of experimentation closely. And given that they seemed so obvious to so many ethicists than others, it made me wonder what the logic of the approach was. And I think it's a phenomenon that we're seeing more and more in science, which is that there's a desire to create experimental groups that are as far apart as possible in order to have a better chance of showing an effect of the experimental hypothesis.

0:05:36.2 CP: In other words, in this case, if you have the idea that a high dose of vitamin D might reduce the number of asthma attacks children might have and you test it against children who are at a normal level of vitamin D in their blood serum, well, you might not get the anticipated dramatic effect that you think you're after. But if you maintain those kids on a placebo and you know that they're deficient in vitamin D, you might be more likely to show an experimental effect.

0:06:05.6 SC: The other option would be to go for a larger sample size, so that you could show a subtle difference and make sure that it was meaningful.
0:06:12.9 CP: Yes. That would be another way of doing it, but they had tremendous problems recruiting for this experiment.

0:06:18.0 SC: And then the trial ended early.

0:06:20.6 CP: The experiment was discontinued for futility. After a certain period of time they realized, "No, vitamin D supplements are not reducing the number of severe asthma attacks that the kids are having." So the data safety board said, "Look, we've gotta stop this experiment because we know it's not working." However, notwithstanding that decision, they continued the experiment, including continuing kids deficient in vitamin D on a placebo for another six months...

0:06:50.2 SC: Wow.

0:06:50.8 CP: To do what they called an 'orderly close-out'. It's a little unclear what the reason for that was.

0:06:56.9 SC: We know vitamin D deficiency is not good for kids. Were the parents or the guardians made aware that their children were not going to be getting these vitamins, even if they were deficient?

0:07:07.9 CP: The fundamental concern that ethicists, the whistleblower and experts in trial design all shared was that the disclosures to parents of these children were highly inaccurate. The fundamental problem with these disclosures were that the main risk of vitamin D deficiency in this age group is bone fractures. And instead, the risk that was stated was rickets. Now, rickets is a disease that can cause fractures and can cause lower bone density, but in this country, it's a disease of infants and very, very small children, much younger children than those that were in the age group tested in this study.

0:07:52.2 SC: So there was a risk of bone fractures. Did they see that happen among the kids that got the placebo?

0:08:00.1 CP: This is an aspect of the study that is particularly worrisome to the folks who have been criticizing it. What they found was that there were nine bone fractures during the course of the study, a minimum of nine, I should say, because that estimate was given before the study had completely ended. And these nine bone fractures were vastly higher than the expected rate of fractures, even among asthmatic kids who are taking inhaled corticosteroids and sometimes oral corticosteroids, which is how these kids were treated in the experiment. So you have a very high rate of bone fractures, however, none of those fractures was disclosed in the journal article describing the study, and none was described in ClinicalTrials.gov, which is the government repository for clinical trial results. Normally, these kinds of serious adverse events would be a required component of any sort of study or clinical trial report.

0:09:04.2 SC: Because these fractures weren't described in any detail, we don't know if more occurred in the placebo group or the high vitamin D group. We don't know if they're from low
impact walking or high impact sports, could something else be happening here?

0:09:18.2 CP: The biggest likelihood is that the low-dose may have had a deleterious effect.

0:09:24.7 SC: The no-dose.

0:09:24.9 CP: Yes. The no-dose had a deleterious effect on bones, that's the major risk in this age group.

0:09:32.1 SC: But we shouldn't be guessing.

0:09:33.4 CP: We shouldn't be. And the only reason we know about the fractures is that a member of Congress who was concerned about the experiment, had been contacted by the whistleblower in this case, he requested that information from NIH and they released the number of fractures partway through the experiment to him, but never recorded it in any other way. I know about it because I have received that correspondence and was able to go over it.

0:10:00.1 SC: Yeah, we should mention that this reporting is based on a lot of research and requests of records and talking to experts and talking to whistleblowers to really try to figure out what happened in this trial. Kind of expanding the picture here, this is an egregious example of comparing experimental treatment with placebo, but this is a trend where people are turning away from comparing standard of care to new intervention, and saying, "Why don't we do placebo versus new intervention?" Why do you think that that's happening?

0:10:32.0 CP: This Vit-D-Kids Trial could easily be dismissed as just a controversial outlier, but these kinds of studies are more and more popular among researchers and the funding is growing dramatically. When I say these kinds of studies, what I mean is that studies that reject groups receiving usual care, in other words, those best current practice treatments used by doctors, in a way of looking for the most compelling results, the most dramatic differences between their experimental groups, researchers sometimes favored this sharply divergent treatment arms in a trial. But the extreme comparisons that can occur in those cases mean that they can't necessarily learn if a new treatment is better or worse than usual care because they're not testing against usual care.

0:11:22.9 SC: We've been focusing on the vitamin D trial, but similar designs have been used for conditions like sepsis. This is a body-wide response to infection that is life-threatening.

0:11:33.9 CP: The goal of this research was to determine the optimal treatment for people who have life-threatening septic shock and that's a kind of terrible symptom associated with that septic infection. So the critics said both treatments deviated sharply from the usual care and that can vary based on their conditions, therefore, they increased the patient's risk of death by assigning them to one of these groups without considering what the usual care would have been for each patient.

0:12:05.0 SC: So neither one received usual care. Is that even clinically meaningful?

0:12:09.3 CP: The sponsors and the supporters of this experiment took the position that both
treatments could represent usual care, and therefore, it was fine to randomly assign patients to one or the other. Of course, the critics say that you have to individualize care in these sorts of circumstances. Now, it really raises an important issue associated with this debate about the design of experiments that lack a "usual care group." And that debate involves the difficulty at times of getting an agreement within the medical community of what constitutes usual care. Sometimes it varies from location to location and sometimes, there's just simple disagreements about what should be done in a particular medical ailment. Now, critics, they would say that it is possible to determine usual care, you just have to do the work, the research, even sometimes surveys in order to learn what the best practices are.

0:13:11.0 SC: You mentioned a little bit earlier that more money was being spent on these comparative trials. So where is that money coming from?

0:13:18.9 CP: About 10 years ago, Congress created a new research funder called the Patient-Centered Outcomes Research Institute, and this organization has been endowed with enormous amount of funds to dramatically increase the attention being paid to comparative effectiveness research, in other words, studies that look at two different methods of doing the same thing and seeing which one is better. So the funding has increased dramatically in the last decade into the multiple billions of dollars being spent on these kinds of experiments, more than an order of magnitude, more than was being spent the decade earlier.

0:13:57.4 SC: So you know that there's this increase in funding from PCORI for this type of research, are you seeing many more studies being done this way?

0:14:06.1 CP: So there's two ways in which I quantify the issue of how often usual care comparator groups are included in clinical trials. One was based on a study done by an expert at an NIH who is considered one of the premier experts in trial design in the country, and what he and his colleagues did is looked at three major medical journals and they looked at the critical care studies that were written up in those journals over a period of a year, and they found that in up to half the cases, depending on the journal, that there was no usual care comparator group in the trials.

0:14:47.3 CP: Consequently, in the view of this expert, there simply was no way in which the experimenters could conclude the medical validity of the results of their trial 'cause they never compared it to what doctors normally do in those circumstances. Another way of looking at it is PCORI data, which is much broader. PCORI funds hundreds of studies in all different realms and only less than 30% of their trials involve the use of usual care comparators. So what we can say overall is that in comparative effectiveness research, a minority of funded trials use usual care comparators.

0:15:28.1 SC: There seems to be this big division; two camps, one camp saying it's risky to patients and not clinically meaningful to give patients placebo instead of standard-of-care in a trial. The other camp says that there's not often a standard that we can use, and we wanna be able to really see the effects here on these trials. So how do you see this resolving?

0:15:51.3 CP: I think it's a big question mark whether this set of concerns about the lack of usual
care groups in these experiments will end up being resolved in the immediate future. What we do know is that time after time, this vitamin D case, being the latest example, there's really, really strong concerns about how these cases are potentially distorting medical practice and critics are raising more and more examples, sometimes resulting in changes in how the experiments are done, but often not. And even the top officials of the NIH agree that this debate is critically important for the future of understanding how best to conduct comparative research.

0:16:38.8 SC: Okay, thank you so much, Charlie.

0:16:40.8 CP: Okay, thank you, Sarah.

0:16:40.9 SC: Charles Piller is an investigative journalist for Science. You can find the link to the article we discussed at sciencemag.org/podcast. Stay tuned for a chat with researcher Birhanu Eshete about the dangers of releasing machine learning outputs on the internet.

[music]

0:17:05.0 Joel Goldberg: Now we'll speak with Birhanu Eshete, Assistant Professor of Computer and Information Science at University of Michigan-Dearborn. Eshete's new article in Science calls attention to the potential pitfalls of the machine learning revolution and why it's critical that policymakers act to address them before it's too late. Hi, Birhanu.

0:17:26.5 Birhanu Eshete: Hey Joel, thanks for having me.

0:17:28.2 JG: At this point, machine learning is everywhere; self-driving cars, medical procedures, intellectual property, even national defense. These applications pose huge advantages but also major risks. Can you explain the risks associated with machine learning?

0:17:45.3 BE: You have this very shiny story behind machine learning solving almost every problem, but the deeper you dig into how machine learning models work and how these models are trained for important tasks, such as helping out a medical doctor to make a decision on a picture of a tumor of a patient is cancerous or not... So in that case, it's obviously super risky.

0:18:08.3 JG: What if some bad actor or adversary, as you call it in your piece, attacks machine learning technology?

0:18:15.1 BE: Until let's say a decade ago, the focus of the machine learning community was just how to make this machine learning community was just how to make this machine learning models more accurate. So they really thought about, "What if I deploy these models in adverse scenarios or situations where, for example, I have a prediction API that I exposed to a web interface, so anybody from anywhere can send an input to a model, and then get predictions?" So in such scenarios, seeing as anybody from anywhere can send an input, you don't have any control as someone who is exposing or serving this machine learning model.

0:18:52.2 BE: As an adversary, I might modify a benign or a legitimate input, and then trick them on it, so by tricking, I mean, the model says, "This input belongs to this group X at one time" and
then after a slight modification, that is, for example, in the case of images imperceptible to human beings, the best example is imagine a self-driving car, which is powered by a machine learning model. And one of the components of the model basically looking at camera feeds and detecting or identifying traffic signs on the road. And based on that, the car should take actions as if a real driver would do.

0:19:27.6 BE: So suppose you show it a stop sign image and then it will classify it as a stop sign, and then the self-driving car will take what you should take as a human, which is basically stopping. And then after a while, you just put very tiny stickers, very little ones that won't even cover the stop text on the sign, and then all of a sudden, the same model that say the image was a stop sign would say this is a yield sign or a speed limit sign. So in that case, the next action of the self-driving car is it's not gonna stop. So obviously, that's a very safety critical situation. This example about the stop sign is specifically called adversarial example, or in more formal terms, we call it evasion attack. So basically, you evade or you trick or you fool the model.

0:20:19.2 JG: Okay, that's an evasion attack. What other attacks can occur?

0:20:22.8 BE: The other attacks that I talk about in the article are poisoning. So poisoning is, before the model is deployed, the adversary might poison the training set, so that the model, once it is trained, would be biased towards some intentions of the adversary. For example, it favors a specific prediction that the adversary wants. So that's the impact of poisoning. The other attack is basically once this model are now being exposed as machine learning data service, which means you can pay per prediction, like you pay per use for different utilities. In this scenarios, by just interacting with the models, adversaries can extract the models or approximate the model, and get more or less functionally equivalent model.

0:21:12.0 BE: We've seen many examples where the big machine learning models that were exposed to APIs from the big companies like Google and so on, have been shown to be vulnerable to extraction or approximation impacts. The risk with such kind of attacks is that when the model is trained on, let's say, intellectual property, basically, you are stealing the intellectual property of somebody else. Even worse, suppose the model is trained on, let's say, national security secrets or that kind of very sensitive data, then again, you're stealing some secrets from a government or a state.

0:21:49.0 BE: And then the other attack that is privacy-motivated is what we call membership inference. And in this attack, the idea is, again, the adversary interacts with the model by just giving it an input and observing the output. But the idea is by just looking at how the model behaves to the members of the training set versus the non-members, the adversary can make a probabilistic estimation of whether a given target is in the training set or not. So what is the implication of this in real life? Suppose you have a dataset from a hospital of patients that you train the model on. If an adversary, by just interacting with the model, can approximate whether my data has been used to train the model, this is clearly a privacy breach, right?

0:22:37.1 JG: One thing that I found less specific in the article was who these adversaries are.
Like any other security and privacy threats, there are multiple actors, but we can point out some realistic actors that would be of interest. Suppose you are a company, you have a dataset that you have collected, let's say for a decade, and your competitive advantage in the market, it comes from that data set, because you are doing predictive analytics on that using machine learning models. But because you have to also interact with customers and provide service through this machine learning model via predictions online, you have to expose it as a public internet-facing service.

In such a scenario, you're essentially endangering your intellectual property, and if somebody else, a competitor perhaps, is able to do the kind of attack, they will have nearly the same competitive advantage. You can cast the same scenario to government. So if a government is providing services that are powered by machine learning models, any government has adversaries who are motivated by politics, financial, and other geopolitical issues, they might also learn a lot of insights by just interacting with this models.

And look at, for example, hospitals, if they want to collaborate on something, let's say, predicting whether somebody has a disease or not, but they don't have to share the data because of regulations like HIPAA, then they might end up exposing their prediction APIs. But if somebody is able to identify a person, an individual, this could be someone who is blackmailing a person or who is trying to ask money just because they have some secret about people. So this kind of threats are realistic.

What can be done to address the blind spots in machine learning?

We are looking at solutions where, in the first place, we have to think about how machine learning models are trained because it seems pretty fundamental that the way we have training machine learning models is making them somehow exposed to this kind of attacks we're talking about. And the other is the assumptions that we had in machine learning models. In real life, you cannot really control the distribution from which inputs will come. So when you test a model in the lab or in a given enterprise, you have control over the training data, you have control over the testing data and you know, you can outline the accuracy of the model and you might say, "We're good to go and deploy", but once you do that and expose it to the rest of the world, you can't force an adversary or even a benign user to draw examples or inputs from the same distribution on which the model was trained.

Machine learning has a lot of problems, but that is not to discourage them. A lot of segments of the community, the scientific community, the public sectors community and so on, they're looking at the issue from different angles: Technical, legislative and so on, but they should be cautious. When somebody comes and pitches a machine learning model as a solution for some important problem, they have to step back and ask questions like, "Okay, what if I deploy this model and somebody steals the model? What if this model is biased towards some segment of my customers?"

There's already a bias on the ground in people, for example, screening the job applicants based on historical data and historical data, usually, has some biases, for example in
STEM fields, like in computer science, engineering, and so on. Others' data is biased more towards male applicants versus females, therefore, it entails a huge risk if you simply use that data. So they have to ask this fundamental questions that I try to outline in the article before they deploy.

0:26:37.4 JG: Where do we stand in terms of shoring up these issues?

0:26:39.9 BE: The timing now is somehow good because we have not yet seen a widespread pitfall of machine learning. We have seen examples here and there, and we have seen people impacted by this. I'm not undermining that, by any means, even one person hurt by this kind of behaviors or misuses of machine learning models is important. But still, I believe that we are at this juncture in the adoption of machine learning models that we have a huge window of opportunity to think about, "Okay, what is wrong with the way we are using machine learning models and who should do what?"

0:27:17.0 JG: Yes, who can fix these problems? How can companies and governments and others protect against these kinds of attacks?

0:27:25.0 BE: Of course, academia and industry are the big players here because academia looks at the problem theoretically and tries to formulate some solutions and test them, but then the big players, like the big companies, which have these big resources for training huge machine learning models that are used for image recognition, language models and so on, they do have the responsibility as well, because in a way, at this point, the data is so concentrated among this big players or bigger companies that have a lot of power, in terms of human resources, compute power and others. But the puzzle won't be complete if we do not involve the public sector. Policymakers are the legislative body of the government because it requires more than technical solutions, it requires regulations. In 5-10 years, it's hard to tell, but we'll definitely be in a better place.

0:28:25.3 JG: Thanks, Birhanu.

0:28:27.6 BE: Thanks for having me, Joel.

0:28:28.5 JG: Birhanu Eshete is an Assistant Professor of Computer and Information Science at University of Michigan-Dearborn. To find a link to his article, as well as other research from this week's issue, please visit sciencemag.org/podcast.

0:28:42.0 SC: And that concludes this edition of the Science Podcast. If you have any comments or suggestions for the show, write to us at sciencepodcast@aaas.org. You can listen to the show on the Science website at sciencemag.org/podcast. On the site, you'll find links to the research and news discussed in the episode. And of course, you can subscribe anywhere you get your podcast. This show was edited and produced by Sarah Crespi with production help from Podigy, Meagan Cantwell and Joel Goldberg. Special thanks to intern Claire Hogan. Transcripts are by Scribie and Jeffrey Cook composed the music. On behalf of Science Magazine and its publisher, AAAS, thanks for joining us.