

Harvard Catalyst | Georg Gerber Episode V1

From the campus of Harvard Medical School, this is *Think Research*, a podcast devoted to the stories behind clinical research. I'm Oby.

And I'm Brendan. And we are your hosts. *Think Research* is brought to you by Harvard Catalyst, Harvard University's Clinical and Translational Science Center.

And by NCATS, the National Center for Advancing Translational Sciences.

[MUSIC PLAYING]

C. difficile is the most common cause of infections in hospitals. Of people who become infected, 25 percent recur, which can be life threatening for people whose health is already compromised. *C. difficile* is well understood clinically. But how to predict, treat, and prevent its recurrence is unclear.

Dr. Georg Gerber is using sequencing of the microbiome to better identify and predict recurrence of *C. difficile* with new data collection techniques and computational methods. Dr. Georg Gerber is co-director of the Massachusetts Host Microbiome Center, assistant professor of pathology at Harvard Medical School, and associate pathologist at the Brigham and Women's Hospital Center for Advanced Molecular Diagnostics. Dr. Gerber, thank you for joining us. Welcome to the podcast.

Yeah, hi.

So you're working on a study trying to predict who will get a recurring *C. difficile* infection. Can you tell us who your patients are? And what are the aims of the study?

Yeah, sure. So just to start with the aims of this study, as we had talked about at some point, *C. diff* is the most common cause of infection in the hospital. About 25% of people that get the initial infection, they will recur with it. And so we're very interested in can we predict who's going to recur, really ultimately to prioritize them for treatment.

So for this initial study, the patient population, it's primarily Brigham and Women's Hospital. But we also have some with Faulkner Hospital and Newton Wellesley Hospital for recruiting patients. And the goal-- this is a relatively small study. So the goal is to have as clean as possible patient populations. So these are people who are relatively healthy other than having the initial *C. diff* infection. So we exclude things like inflammatory bowel disease, immunosuppression due to cancer, and that type of thing.

And how is the study being funded?

So this is from a program at the Brigham called Precision Medicine, which is a donor-funded initiative. And it funds a variety of things. But part of it is they funded some primary research projects that are focused around this idea of precision medicine. And this really falls into that category is, can we look at people who might recur? And can we actually predict that and then give them a more precise and tailored treatment?

OK. Could you talk a little bit more about how this-- what you just mentioned, this fits into that precision medicine idea, how does this study-- and maybe if you could give us a little background on what this idea of precision medicine is. How does this study fit into that?

Yeah, yeah. Now, I mean, that's a great question. I think this is something-- it's a term that's come up in the last, I guess, around five years. And people are still working out exactly what it means. I think it means different things to different groups.

But the idea is that you have a population. And you may be trying to predict what's going to happen to that population to pre-anticipate a preventative way. Or you may have someone coming in with a disease. But you want to, in some ways, manage and predict that process is a lot of what I see going on in precision medicine.

And so this project really fits into this idea that, OK, we have-- if you take 100 people and they have C. diff, we know, on average, 25 of them are going to recur. But we don't know who. And it turns out there are actually fairly effective treatments for recurrent C. diff. There are new antibiotics.

There's also fecal transplant, which we'll probably talk about subsequently, and a bunch of other modalities coming on. But they're expensive. They have side effects. I mean, you don't want to treat those 75 people who aren't going to get it. And so the precision part comes in. Can I identify this 25 people, give them what they need so they don't recur?

And you're trying to do that when they first present in the hospital rather than having them come back with a recurrence? Like, you can prevent that recurrence if you feel that they have a high likelihood of recurring?

That's right. So you want to prevent the recurrence. And part of the question of the study is how soon can we predict who's going to recur. It may not be right when they come into treatment. We may have to have them come back in a week, two weeks. That's part of what we're evaluating. But it's very costly, and also a huge amount of morbidity if someone actually does recur with the infection. So you want to prevent that happening right at the onset if you can.

And so C. diff is a pretty serious infection. Maybe you could tell us a little bit about C. diff. And you mentioned the morbidity when somebody recurs, meaning the severity of the infection. Maybe you could talk about the initial infection, why that's a problem, and then why the recurrence is an even bigger problem.

Yeah, sure. So C. diff is an interesting phenomena. Actually, since antibiotics first came online, which is about in the '50s, it was observed that people that had a lot of antibiotics would have these diarrheal episodes. They didn't know why. And that was really ferreted out around the 1980s with implicating C. diff as the primary cause of this.

And so the usual course is you take antibiotics for something. Classically, it might be you have a tooth abscess or get a dental implant, something like that, where you get clindamyci is one of the big offenders. And then you come down with this diarrheal infection. And it's not like just garden variety diarrhea. It's very bad diarrhea.

Patients typically are having to sit on the toilet for many, many, many times a day, hours a day, very, very debilitating disease. If you're a healthy adult, you're probably going to come out of this OK, assuming you don't recur. But for elderly population, people who are immunocompromised, it can be much more severe.

And so certainly, there is mortality. There's about 15,000 deaths a year from C. diff. And you can also get complications. I mean, you can literally can destroy your colon. There's a phenomena where the colon becomes so inflamed and so damaged that the patient can either die or you have to take it out. So you can get these very severe complications.

So that's the primary episode, which is bad enough. But yeah, in people who recur-- and that's-- we're still-- other research projects, we're trying to understand more about the biology of that. But we don't completely understand why people recur and the whole phenomena. But the bottom line is if you recur once, you're more likely to keep recurring. So there's something going on, probably in the patients, but also with the C. diff strain.

The impact for the patient is huge. So there are people that will have this disease for months and months, even years. And as I said, it's a thing where they're sitting on the toilet most of the day, massive amounts of weight loss. You have vitamin deficiencies, all these things that go along with having this just chronic debilitating diarrhea.

So let's get back to the study itself. And let's talk about the kind of data you're collecting. And you're also looking at the microbiome and using microbiome sequencing. So maybe you can tell us what kind of data you're collecting, how you're collecting it, and how you are trying to use that data to identify-- or to target these issues.

Yeah, so that's a great idea. I mean, we'll start with the gross part of the study, literally and figuratively. So what we do is we try and identify patients as quickly as possible. We have an interesting system at the hospital where we get alerted to when someone tests in, so a lab tests positive for *C. diff*. And then we run over and try and consent those patients.

And the goal is to follow them for eight weeks. So we try and get a stool sample every week from each patient in the study. As you said, we're working with microbiome. Or we want to get microbiome data. There's a whole protocol for getting that. But basically, what it amounts to is that we need to get the sample frozen. And then we have a system for patients shipping the samples to us.

From the microbiome perspective, the initial part of this study is to ask which microbes may be present that could be related to the recurrence risk. And so for that, as you mentioned, we use sequencing, which I can go into a little bit more about that modality. The next data the type that we're collecting-- and we just got additional funding from the Precision Medicine program as well as Catalyst-- is to look at the metabolomic profile, which is another modality that's very, very interesting, particularly for *C. diff* infection.

So tell us a little bit more about the microbiome sequencing and what-- we've talked a bit about on the podcast about microbiome research. But how are you using this? Or what kind of work are you doing around this?

Yeah, so in the microbiome field, sequencing and looking at nucleic acids has been a huge enabling piece of it. There's certainly many of the ideas in microbes have actually been around since, again, probably since the 1950s, even earlier. But really, what enabled that was high throughput sequencing and the cost efficiency with that.

So there's really two main ways that people sequence. Right now, the first would be called shotgun metagenomics, where you take all of the microbial DNA in that sample, just break it up, and sequence it. The second approach is looking at essentially a marker gene, which is the 16S ribosomal RNA gene. And it's a marker, if you will that, is highly conserved. So it's present in every bacterial species. But you get enough diversity into it that you can identify, more or less, what's there.

The gap with that is the resolution of what you can identify is somewhat limited. So you may be able to get down to the genus level typically, but not necessarily species and certainly not the particular strains of bugs that are there. Whereas with the metagenomic approach, you can go deeper. But it's much, much more expensive right now. So for our study, we did use 16S, in part because we're looking at patients over time. And so we need multiple samples. We just did not have the funding to do this shotgun metagenomics.

OK. So what you're talking about is there is a trade off between quality and cost. And you're-- so you have a lower cost, but also somewhat lower quality. So how does that affect what you're able to determine from this study?

Yeah. So one of the ways we designed this study was to look longitudinally, which is quite important, because there's a lot of temporal variability in the microbiome. You also have a process as the patient is-- they're coming they're treated with antibiotics for the infections. So they're coming off antibiotics. Their microbiome isn't static. So it's important to look at that process and change over time as well.

So to my mind, the longitudinal studies give you a lot more information than just looking at a slice in time. But with that comes the question of multiple samples. And so it drives up the cost. So from my perspective, it's worth it to go with the 16S, which, in some ways, is lower resolution, because we have the longitudinal information. And in part, if we see something in there, we see it multiple times or can see it multiple times. That gives us back some of that resolution. It also gives us additional information.

So could you give me an example of what you might find based on the techniques that you're using? You talked about being able to identify genus, but maybe not species. So what would you be able to find that would lead you in a direction of being able to make a determination about what role certain bacteria play in the recurrence?

Yeah. So it's interesting with the resolution, because it's not a uniform thing. So there are some areas in the phylogenetic when you're looking at microbes where they're quite different from other things. And you actually can cleanly get to the species level. There's others where it's going to be somewhere in between, and others where it's going to be very difficult, because you have many things that are very close.

And so part of it has to do with diversity in this particular gene we're looking at. Part of it has to do with diversity in the tree of life. So interestingly, some of the organisms we care most about we can get a bit better resolution on, because they're in these more isolated areas in the phylogenetic tree. So there's a fair amount-- actually, we can, in this particular, study get it out of 16S. And that's why we're more confident in using this technology.

OK. So once you get-- once you start going down the branches of the tree, this branch only leads to three or four different sub-branches. So you're pretty confident that those are what you're looking at.

Right, because there's some areas where you just you don't have that much diversity in the human gut. They may be elsewhere in the environment everywhere. But you know in the gut it may be two or three species. And you actually can get down to it with 16S.

So is there anything else about the sequencing or the data collection that we should talk about?

Well, it's probably worth talking with the metabolomic data.

Yes. You want to go into that now?

Yeah. So I think, traditionally, as I said, a lot of the microbiome has been based on sequencing data. That is interesting, because you can identify what's there. You can also potentially identify genes. But this just gives you names of things. And the question becomes what are their end activities that might be affecting whatever they're doing.

And so with microbes, a lot of their life is eating things and excreting things. They're relatively simple on a certain level. And so looking at their metabolic outputs can be incredibly informative. And particularly in the case of *C. diff*, where-- so it's been recognized for a long time, as I said, that when you have an antibiotic treatment, this makes you very susceptible to the infection.

Similarly, there's been these recent studies where you put back in-- so you give these fecal transplants. And you actually get very efficacious at treating *C. diff*. And so clearly, which microbes are there are playing a big role. And there's a lot of hypothesis that the way those microbes are working is somehow competing with *C. diff* for something. And so a decent hypothesis is that it's metabolic competition, because it's a lot of what microbes are doing. So the question is somehow, how are they altering the gut metabolic environment to make it less hospitable for *C. diff*?

So are they eating all the food so *C. diff* can't get any food? Or am I on the right track there?

Yeah. So one of this, yes, would be nutritional is, OK, there's some carbon source that they can eat faster than *C. diff*. Maybe it's also a vitamin that they're consuming that *C. diff* needs. They're also, though-- interestingly, microbes use metabolites in a sense for signaling in a sense, where they will sense a metabolite. And this makes them do something else differently. So *C. diff* is a pathogen, because it makes a toxin that kills off epithelial cells.

Epithelial cells are the cells that line the organs like the colon.

Yeah. So in the case of the gut, what's happening is literally *C. diff* makes this toxin. And then your cells in the gut are being killed off by it. That's why you get this-- eventually, you can get this really bad diarrhea. And so one hypothesis-- and this has been observed to some extent-- is how *C. diff* is active metabolically affects its ability to produce the toxin.

If you think of it-- a very interesting fact-- a large number of people are walking around colonized with *C. diff* and they don't get sick. So there's something going on that's just changing that environment. And if you think about the microbes as what their lifestyle is, they become pathogenic. They'll make a toxin which is very metabolically expensive for them. But they'll do that when they're stressed.

And so they're in an environment where they're sensing, I don't have food, I don't have what I need. So I need to elaborate this toxin, basically get food and get what I need. And so the idea is the other organisms are doing something metabolically changing that environment, which is making *C. diff* behave like a pathogen, secrete toxin, and be bad, or just sit there and hang out and not do anything.

Interesting. So most of the time, *C. diff* is not doing anything? It's being held in check. And it's only when-- but we don't know why it gets stressed or why it starts producing the toxin?

Yeah. So I mean, that's-- we actually have another project that's also been funded out of the Precision Medicine, which involves-- it's an animal model. And we have gotten it down to the metabolites and directly what's going on. But in the human study, we don't know. That is still the operative question is what the command cells are doing to make *C. diff* not toxigenic or, when they go away, why it decides it needs to be a pathogen.

Can you talk about how new computational methods can help to accelerate these really important discoveries?

Yeah. So I mean, broadly, a lot of the next generation sequencing revolution-- this goes way beyond microbiome, but obviously looking at human genome, cancer, and things like that-- have been completely dependent on computational techniques and technologies. All of our sequencing techniques that are really in regular use are dependent on short reads. And so this gives us a fragmented picture of what's going on. And you have to use the computer to go back in there and reassemble what's going on.

So on the core bioinformatics front, that's been absolutely essential for all these high-throughput technologies. The next phase we're starting to see broadly in the biomedical arena is machine learning, and the idea being that you can start to collect now with high-throughput technologies lots and lots of samples. And you're asking a question, OK, can we predict something from a sample?

So in our study, we're trying to predict *C. diff* recurrence. But you might also be-- you want to predict someone's going to have cancer, what their life span is, all these things. And can you take all this information that's very rich, very complicated and get signal out of it? And that's where machine learning is starting to come in in the biomedical area, which we've had. It's been around in other fields, but relatively new in biomedicine.

And so machine learning can pick up the small differences that wouldn't be detectable otherwise, or making sense out of large amounts of data, looking for patterns?

Yeah. I view it a lot as the latter is what it's good at is where you have a very complex type of input. So it could be lots and lots of samples or lots and lots of things we're measuring, like genes or microbes. But there's also complexity to it. So these things could be interacting in complicated ways. So you might need 10 times a and 300 times b and them interacting in some way. And so there's this complexity in figuring out how the parts interact. But you can't just sit down and look at a printout of this and figure that out.

Is this informing treatment? Or does this go beyond treatment to prophylaxis? Like, OK, we think you're going to get this. So start doing this now and you won't get it.

Yeah. So it's an interesting question. And I think that people would like to do both. And there's very compelling reasons to do both. Part of where things have been driven have to do with factors that have nothing to do with science. And so how things are funded, and preventative medicine in general, what the interest is in this versus a treatment.

So I think the first tranche of things-- there has been a lot of emphasis put on treatment. And so in the cancer space, which is the other thing I do as a pathologist, we are now using next generation sequencing to say which treatment we should use on a patient that matches the mutation to the treatment modality of the tumor. What we'd love to get to, though, is you're able to look at the sample, predict who's going to get disease a year out, five years out, whatever, and then do something so they never get it.

I think, yeah, everybody would like to take a blood test and then get a list of, OK, you're going to get this disease in this year, and then-- but that's-- I mean, that's far off, wouldn't you say? It sounds like science fiction almost.

I think there's aspects of it that are science fiction. And there's some diseases that are very, very complicated, because you have environmental components. You have genetic components. You have a lot of just random components. But not everything. I mean, there's some where we have pretty clear causation. Or we could probably find pretty clear causation and make those predictions pretty accurately and pretty far ahead of time. Yeah, it's going to depend in part on the disease.

But the other piece of it-- and I think this is where people are excited in the machine learning world-- is if you could collect enough information on a patient, could that truly be enough to predict even a complex disease, like say onset who's going to get type 2 diabetes and when? Could we actually figure that out if we're monitoring someone's diet, if we're looking at their daily body weight, if we're looking at physiological areas, blood labs, et cetera? And I think no one really knows at this stage. But that is the dream.

The interesting thing with machine learning right now is that a lot of the best methods, the ones perform well-- if you look at how Siri works for recognizing voice or techniques for Google that can do image recognition, they use so-called black box methods, meaning that there's a very complicated mathematical function that you give it lots of data and it learns that function. But it doesn't give me the ability to look inside that and understand how it's working. It's just way too complicated and unstructured. This might be reasonable for those type of applications. But in biomedical applications, you probably want to know why things are happening.

Right. And for research and-- a lot of research is based on peer review and everybody checking the methods and seeing and replicating it. So you'd want to be able to take that function, replicate it in a different population, and see if it still holds true.

True. But I mean, that you could conceivably do empirically, if the goal were just, OK, I'm going to reproduce the result. But the other question of discovery is, have I learned some general principles? And I agree with you that those two things are not independent, because if it's a general principle, it's probably going to be much more likely to reproduce another population, right?

So I think those two things are arguments against why you don't want the black box. The other thing is, sociologically, if you're thinking of this as going to become a diagnostic that physicians might use, physicians are pretty allergic to just, I put it into a black box, got an answer. I mean, they want to understand.

Why you got that answer.

And the patient wants to understand. Can I explain why and how we got the answer? So those are all pieces why this black box mentality isn't exactly what we wanted in biomedical research. So part of what the Catalyst program is we're developing new algorithms that have the same power as these black box methods, but are human interpretable.

And so they literally produce rules that say, if this particular microbe or metabolite is above a certain concentration over a certain time window, and perhaps you need another one that's doing something else in combination, that increases your risk of recurrence by x percent. So you can actually look at what the predictor is and figure out how it's making the prediction.

And so how is that-- I guess, how is that better than-- you explained the sociological reasoning why you'd want to get away from the black box. But how else is that advantageous to a physician or a researcher?

Well, from the research perspective, it goes back to one of your earlier questions, which, OK, I have predictors of who's going to recur. What does it tell me about the microbes? Would it tell me what's going on? And from our perspective, we ultimately want to figure out scientifically what microbes are inhibiting, how they're inhibiting, but ultimately from that improves the treatment.

So we don't just want something where I throw a bunch of variables into a black box, it says you have some percent of chance of getting C. diff again. I'd actually like to know what the causative factors are. And so that's really the piece that I'm most excited about with this is, with enough data, it can eventually get us-- enough data and also doing our animal experiments over modalities can get us to causality in the system.

And then from there, you can develop a treatment.

Right.

So how will this research contribute to treatment for recurrent C. diff?

One goal with the research really is just to predict who's going to recur. And so you could view it as biomarkers, where we might-- they may be metabolites. They may also be microbes where we know their nucleic sequences, nucleic acid sequences. And you can put those in. And there's some predict that would say, OK, this patient is this percent chance going to recur. This other one is unlikely to recur.

The other piece of it, though, is looking into what those predictors are. And that's very much enabled by our work with Catalyst, with developing these machine learning tools where we can get human interpretability. And so the idea is, OK, if we can discover from this which microbes and which metabolites are predictive of recurrence, that enables us to trace back and figure out, OK, those microbes, if present, may actually enable you to not get recurrent disease.

So from the treatment end with C. diff, there's a few different modalities people are looking at. One of them I mentioned is this fecal microbiome transplant, where you just take a stool from a healthy donor. And you basically put this-- now, it's usually in pill form. The patient takes it. And the idea is that it's somehow going to repopulate their gut with something.

The trouble with that is we don't specifically know which bugs are involved. There also are safety risks. We're taking an uncharacterized biologic and putting into patients. And there's also efficacy questions. The initial studies showed this was very efficacious. But those were in pretty restricted populations. And to your point of generalizability and reproducibility is people are going out to more diverse populations. This may not have the same cure rate.

And so I think that the dream with the precision therapeutic is to say, I'm going to identify particular microbes, or even metabolites, that are essential for fighting off this C. diff, so it either never takes hold or you know wouldn't even be able to colonize well. And so that's part of the program we have with Precision Medicine on our animal models. We've actually identified specific bacteria that are protective in mice against the C. diff challenge. And we've even gotten it down to which metabolites are protective in this model.

And so then you can develop-- so how do you treat them? How do you then treat the patient with that information?

Well, so the idea is-- and this is what we're unsure of is, is there going to be a universal set of bacteria metabolites that will work for all people? In which case, if it's metabolites, then you figure out-- that's more of a classic pharmaceutical thing, where it's a small molecule. If it's bugs, you would have some means of giving this patient, probably in a pill-- people have been talking about some sort of yogurt formulation. And they would get the bugs to repopulate, recolonize possibly after antibiotics or even preceding this prophylactic and be resistant to the infection.

Dr. Gerber, thank you very much for coming in. It was a pleasure to have this conversation with you.

Yeah. It was great coming here and interesting questions. Yeah, fun to do.

Next time on *Think Research*.

So we knew a great deal about some causes of infertility. But then there is a good proportion of couples who aren't able to get pregnant. And we're really not sure why.

Dr. Caroline Mitchell considers how the vegetable microbiome impacts IVF success rates. Thank you for listening. If you've enjoyed this podcast, please rate us on iTunes and help us spread the word about the amazing research taking place across the Harvard community.

To learn more about the guests on this episode, visit our website catalyst.harvard.edu/thinkresearch.