[Sarah Alger] Welcome to Proto, a podcast that explores the frontiers of medicine, and welcome to Diagnosis, a series about the past, present and future of a medical cornerstone. I'm Sarah Alger.

[Adam Rodman] And I'm Dr. Adam Rodman. As physicians it's hard to conceive of our craft without that moment of diagnosis. Indeed, that single act has been central to really how we discuss what medicine is even in the earliest accounts, dating back thousands of years.

[SA] Attending to sickness in ancient Egypt and Greece echoes in many ways what happens today: A healer examines symptoms, draws on professional and sometimes arcane knowledge, then pinpoints the problem.

[AR] And we can all be very proud that our underlying understanding of disease, the practical epistemology we use, has radically improved over time. Our model of the human body is dramatically more reliable than it was 3,000 years ago, but while it might be more subtle, we can also talk about our evolution in diagnosis, important shifts in the process we use to collect patient symptoms and other clinical information to assign a disease to a patient.

[SA] A brief history of diagnosis, coming up on this episode of the Proto podcast brought to you by Massachusetts General Hospital.

What happens when two physicians disagree about a diagnosis? If it's on a television show, you can bet that one doctor is right and the other is wrong, with the right answer being cleared up by the time the final credits roll. In practice, however, a conflict like this can lead to profound questions. Researchers who study diagnosis have, over the past century, looked at diagnostic disagreement and other issues to rethink core questions about diagnosis itself. Are physicians using consistent criteria every time they look at a patient? Are gold standards of disease, really telling us anything about the body? And what is disease, after all? These questions have been of particular interest to Dr. Adam Rodman, an internal medicine physician at Beth Israel Deaconess Medical Center in Boston, and a passionate medical educator. Dr. Rodman's podcast, Bedside Rounds, explores medical history with a particular focus on how we have understood medicine and the body. In several recent episodes, Bedside Rounds has explored questions around diagnosis, and we are happy to have Dr. Rodman here to share some of that work with us today. Welcome to the Proto podcast.

[AR] Thank you so much, Sarah. I'm so excited to be here and please call me Adam.

[SA] All right. Will do.

A moment ago, I referred to studying physicians who disagree about a diagnosis. Is this a modern idea?

[AR] Absolutely not. Uh, I mean, physicians are ultimately people, people love to disagree. People love to pull rank and hierarchy over each other. So we can even see examples of people disagreeing about diagnoses in Galen, right? So we're talking like, uh, second century CE and people are, Galen is writing about how he disagrees with other doctors. So there's not the sense that diagnosis and disagreement is modern is not true. I think the difference here is that, uh, let's say that I have a disagreement with a Greek physician in Byzantium in you know, the fourth century, the difference would be that I would say I'm a hundred percent right. That person is a hundred percent wrong. They're just an idiot. So up until the modern era, that is what disagreements look like.

[SA] And it seems as though these disagreements took on new urgency in the 20th century. One group in the 1930s and 1940s looked at physicians who recommended a tonsillectomy for their patients. Can you tell us about that?

[AR] So you have this idea called focal infection theory. This idea that, um, disease cancer in particular starts somewhere as an infectious focus in the body. And if you just remove that infectious focus, well, you've cured cancer. And there are like case reports of people removing swollen tonsils leading to not getting cancer. I have no idea how you would actually prove this. So tonsillectomy became by the 1920s became a pretty standard medical procedure. And even to this day, like a lot of my patients who are 60, 70, have had their tonsils removed. It's a much less common procedure now. So in New York City, they actually had a program called The Pathways to Health for school kids.

Again, this is the 1920s, 1930s in the progressive movement. It's a little bit eugenical, but there was this idea that they would identify people who were not on the pathway to health. And that would include things that we would be all about today, like correcting a clubfoot and then also things that are a little more problematic, but one of those things was tonsils and they launched, an audit, I guess, of every single elementary school in this Pathways to Health program to see that kids were getting on the appropriate pathway. And one of the things they looked at was tonsillectomy, and they wanted to make sure that they did a really good job. So they had a panel of five different physicians. And I think they randomly chose about 1,500 kids. Actually, most of them had already had their tonsils out.

So they took the kids who had not had their tonsils out. And there were about 400 of them. They had a doctor look in their tonsils, see if they would benefit from a tonsillectomy, why hadn't they gotten the appropriate level of care? And that doctor looked at these kids and 45% of them, he was like, ah, this is absolutely wrong. You guys need to get a tonsillectomy. Well, the people running the study had some clever ideas. So they took the kids who had not been selected for tonsillectomy and asked another doctor to look at them, a separate doctor, that doctor again looked at these kids and roughly decided that 45% of them needed their tonsils out. They went through this process two more times. So four cycles each time a doctor said roughly 45% of these kids need to have their tonsils out.

I don't think it's any stretch to say that that doesn't make sense, right? Not only are these people disagreeing with each other, they're disagreeing with each other at about the same rate, right? It's uh, the base rate fallacy; just by seeing somebody I know it's gonna be roughly a coin flip of whether I take their tonsils out or not. And this was a big, big deal. In the 1930s and 1940s, this was used as kind of the critical piece of evidence to stop this procedure, stop this unnecessary procedure, though, given that this is medicine, still persisted for a few decades after that.

[SA] I wonder if we can turn to tuberculosis. On your podcast, you discussed the work of biostatistician Jacob Yerushalmy in the 1940s. As I understand it the military wanted to more accurately diagnose TB in its soldiers and at the time TB was based on how a physician interpreted certain images of the lungs.

Can you tell us a little about Yerushalmy the data he was working with and what that work did to change how we think about diagnosis?

[AR] Yeah. I'll call him Yak. That's what he went by and it's much easier to say. Yak is a fascinating character cuz he was already a, I don't wanna say he's a big deal. He's a big deal to people who are really interested in the history of diagnosis already. He did a lot of the preliminary work on linking lung cancer and smoking even though later he would contest that. But World War II comes along and he gets drafted. He was working for the CDC and he gets drafted into the U.S. Army. And the U.S. Army has, they have a problem, right? Tuberculosis is still a big deal, especially in congregate settings. So training camps and bases in Europe and Asia. And they want to know what is the most effective way to screen for for tuberculosis and going into this study, the people who conceive this study, this is a technological feasibility study.

They have four different methods, all versions of x-rays that they want to evaluate to see what's best, right? They're not going into this thinking, oh, we're gonna uncover something fundamental about diagnosis. They just want to know what works best. So in this study they went to two different VA hospitals. These are older world war I vets as well as the people who worked there, which just shows you they're going for convenience samples. And they made every single person in that hospital, the patients and the employees, that ends up being about 400 people total, take a different study of these four different methods of chest x-rays. So they end up, I'm rounding the numbers, but they end up having about 1,600 films total. Um, okay. So Yerushalmy is the guy who now has to sift through all of these and figure out what's the best way to do. I have 1600 films. Um, it is a lot of data. So he devises a pretty clever way. Well, how are you going to tell if somebody has tuberculosis or not? What is the method that we're going to do? So what he decides is that every single, there are five, every single what we would call pulmonologists, they call them phthisists at the time. Pthisis of course is the ancient name for tuberculosis. And it persisted until the 1960s in the U.S., which is kind of cool. Anyway, he has five different phthisists and he makes each of them rate a yes/no, like a binary: Does this person have tuberculosis or not? Then they come back two weeks later and read the same films over again. And then using that data, he figures out what we would today call a reference standard uses group concordance.

And I think it's ending up to be like 10 reads total for each film. Does the majority of them agree that that is tuberculosis, then yes, that is a definitive diagnosis of tuberculosis, and doing that he can calculate the diagnostic accuracy of each of the four methods. In the end, the answer was pretty boring. The answer to this, this study was, eh, they all work. They all work. None is better than the other. There's nothing exciting about it, but Yerushalmy, I wish, oh my goodness, I would love to see what he was thinking at the time, because within a month of this being published, he publishes an incredibly heretical paper with none of the other authors on it for the CDC, with only his own name on the study, reanalyzing this data, coming to some very like troubling conclusions.

[SA] Can you tell us more?

[AR] So Yerushalmy imports these conceptions of diagnostic accuracy called sensitivity and specificity. Uh, should I explain what those are? We talk about these in medicine all the time. We force medical students to make two-by-two tables and everyone hates it. Nobody really learns what it is. Sensitivity effectively is the ability of a test to tell if somebody has a disease, if they actually have the disease. So a very sensitive test will have very few false negatives, meaning that if you're negative, a very sensitive test, if you're negative, you almost certainly don't have the disease. Specificity is the opposite: the ability to tell healthy people that they don't have a disease. So a very specific test is the same, very unlikely to be a false positive. So Yerushalmy decides to reanalyze this data, using the concept of sensitivity and specificity.

I know I'm building this up. It's not very exciting. And again, he reanalyzes this and he discovers that, well, shocker, they're all equally sensitive and specific, but he has this amazing insight here. How do we determine whether or not somebody actually has the disease? Yerushalmy has chosen group concordance, and when he reanalyzes, it, I believe group concordance was, he has five reads. If two of the five are positive for tuberculosis, he's counting that as positive for tuberculosis, but you start to look at how often people agree with each other, and you can see that the readers disagree with each other quite frequently. One of the most troubling is one of the readers here disagreed with himself, his rereads, 50% of the time. This is the very early beginning of the field called psychometrics, which is the study of human perception. And what Yerushalmy is finding is, well, these physicians are very confidently saying yes or no. I mean, maybe they truly do have a disease or don't have a disease, but we have no way of fundamentally knowing whether this is a yes or no. The best that he could do is this group concordance. And he has this like troubling reliability, which is that diagnosis is best given in, in the sense of a probability, of a probability of the person having a disease, because it's being interpreted by flawed humans. Even worse, in order to validate whether a test works or not, you need to compare it against what you might call a gold standard diagnosis. What you would call an EBM. We call it reference standards now. And those reference standards are also subject to a certain amount of unknowability. In this case, his reference standard was just: Do these people agree? and he could look, he could look at individual people and see they disagreed with each other 50% of the time in some cases.

So Yerushalmy actually is very, very careful in his paper. He scopes his findings. He's like, this is only important for screening. He's trying not to offend people. And this is why I wish I knew what he was thinking because he's like, okay, it's fine. This is in screening tests. We get it, like in syphilis testing, similarly, the sensitivity and specificity is going to be important. But, um, Jerzy Neyman, who is a very famous statistician wrote the--so, sorry, I should say the, the, um, the CDC, uh, tuberculosis control division editor dedicated the entire issue to this and Jerzy Neyman's paper, like, throws the football the rest of the way down, because what he concludes is that, look, what Yak is saying here is not just for a screening test for tuberculosis. It applies to every single diagnosis. We cannot think of diagnoses as yes or no.

We can think of them merely as probabilities. And when we think about interpreting tests, we cannot confidently say yes or no. We can only talk about the probability, the sensitivity, the specificity, the test characteristics. There's an even more disturbing realization that Yak has in this, which is the reference standard. You can imagine: His reference standard is looking at a roentgenogram, uh, five people looking at a roentgenogram. Now imagine 30 years later, a CT scan is invented. The CT scan is going to be a much better test, obviously at picking up occult tuberculosis. Well, if your reference standard is these old-fashioned x-rays, it's gonna show up as a false positive, right? Having a reference standard will prevent like more effective diagnostic or make it much harder for more effective diagnostics to come on onto the scene. So he like identifies this kind of fundamental uncertainty and fundamental tension in knowing how diagnostics work all the way back in 1945.

[SA] That seems like a profound insight that maybe gold standards of diagnosis aren't perfect indicators of disease. Can you talk more about that and maybe any more modern examples of that?

[AR] Absolutely. Medical students will often take evidence-based medicine courses and they are taught, as I was taught as a medical student, that sensitivity and specificity are test characteristics. They are unchanging, they are inherent to a test. But it doesn't take many thought experiments to see that that's not true. So there's all sorts of biases that can affect our quote unquote test characteristics. One of the most common is spectrum bias. And I can give some examples about this. So spectrum bias is, you know, you study a single population, but that population might not be the same as the population that you're seeing. So a classic example is something called the Wells criteria that we use. It's a decision tool validated in the emergency medicine world to try to pick up whether somebody has a deep vein thrombosis or a pulmonary embolism, looking at risk factors.

And you might think that you would be able to use it in other settings because it is the test characteristic of the Wells criteria. Well, it turns out it doesn't work like that. It doesn't work in hospitalized inpatients at all. Fortunately, we've studied this and we know it, but it's a perfect example where the spectrum of the patients is different. You can even look at many famous studies that we do to be very confident about our tests. Like I was, there's a paper that I love looking at the workup of anemia, of low blood levels and different serum studies. Well, it was done in a hundred otherwise healthy male veterans at a VA with a confirmatory test. Is that going to be relevant to my mix of patients, which is 65% female, much sicker, is that relevant? And the answer I would assume is no or if not no there’s no reason to think about it.

The final thing that can really bias things is something called workup bias. And we see this all the time.

The problem is that, uh, somebody who has a positive study gets more invasive tests than the person who has a negative study. So again, talking about DVT, the person who has a positive study might get in these studies, like venography squirting veins through the legs, which is painful and has some risk. The healthy person who has no symptoms of the DVT is not getting that. So you don't have a sense of what degree of quote unquote healthy people actually have these findings. And the point of all of these is all of these forms of bias that can affect our test characteristics and doing air quotes, have a tendency to make our tests look like they're more effective than they really are.
[SA] The flip side of that is can you even estimate the percentage of diagnoses that actually are straightforward? Like this person has this germ? Is that just a very tiny fraction, a surprisingly tiny fraction in your opinion?
Dr. Adam Rodman (23:47):

So, so what you're, what you're talking about is nosology, which is the characterization of diseases and there's different ways we can talk about nosology. I'm talking about a very straightforward, does this person have a disease or not? One of my favorite examples is pneumonia. Pneumonia seems about as straightforward as you would think, right? Pneumonia, you have a chest infection, it causes symptoms. Um, you might do a chest x-ray and see it, right? When you start to dig a little deeper though, you come to all sorts of problems. So there's a famous study in patients who have positive chest x-ray everyone in a positive or negative chest x-ray findings. Everyone got a CT. And in about 50% of the people who have a consolidation on the x-ray, there's no consolidation on the CT, but the opposite is true too. In about 50% of the people who have a clear chest x-ray, there's a consolidation on the CT.

So you have this sort of fundamental nosologic question about what pneumonia is. All diagnoses are like this, but they're not all like this to, to the same extent, right? So pneumonia has very, very like fuzzy boundaries. Something like lupus has very, very, very fuzzy boundaries. My goodness. Talk about psychiatric diseases. Those boundaries seem, you know, we have diagnostic criteria. They seem very rigid, but they're voted on by democratic process. Like what sort of like that is a foreign nosology to what we talk about in, in my field, in internal medicine. At the same time, we have diseases that have much stricter boundaries. So if there is a straight-up pathologic diagnosis, so lots of cancer diagnoses are truly a, like, it's, it's a pathological anatomy. They're taking a sample of the tissue and looking at under the microscope. Now, when you start to look at how often different doctors might disagree about something, there's still a disturbing amount of uncertainty in there, but there at least is a more firm pathologic diagnosis. So I would argue that there is a fundamental uncertainty, like there's a fundamental uncertainty in our tests, but there's also a fundamental uncertainty in our diagnoses. And there's not always clear boundaries, but not everything is equally fuzzy. Is that a fair answer?
[SA] Absolutely. Thank you. I wonder if we can change course for a minute. A lot of your work focuses on the shifting nature of what we know the epistemology of medicine in terms of diagnosis specifically, we've also seen some diagnoses come and go because the disease itself stops existing. One example that comes to mind is homosexuality, which was removed as a mental disorder in 1973 because of democratic vote. In that case, the so-called symptoms were simply no longer regarded as a disease. Also proto has covered nostalgia, which was a diagnosis of extreme homesickness that was credited for the deaths of about 70 soldiers in the Civil War. In this case, the deaths might still occur today, but we would probably give them a very different name. Do you have any favorite examples of a diagnosis disappearing or evolving?
[AR] Of course I do, Sarah. So first the, the aside about psychiatry and psychiatric classification of disease: So in both the cases of homosexuality and I think there's been some controversies more recently, such as Asperger's no longer existing and becoming part of autism spectrum disorder. The reason is that still in psychiatry, we, I shouldn't say we, I'm not a psychiatrist, but they use a nosology of symptoms. And how do you develop a nosology of symptoms? Well, as you say, a democratic vote,

right? People come together and decide what symptom complexes might constitute a disease. We actually used to do this in internal medicine. If you go back to the 18th century, it wasn't called internal medicine, just medicine, disease was defined as its symptoms to the point that if your symptoms would change, you would have a different disease. And there's lots of famous examples of symptoms changing a little bit. So doctors would be like, oh, it's not this disease, it's another one. But no understanding that symptoms were manifestation of a disease being underneath. Um, psychiatry still does that today. As an internist, you know, the gut feeling is that it's really weird, but when you actually end up looking at the effects of psychiatric drugs versus the effects of medical drugs, the effect sizes are the same. So it works. And to me, it's this really interesting phenomenon that still exists. Now, my favorite example is purely based in pathological anatomy. It's purely based on a modern understanding of disease. And that's how we've changed to think about kidney disease. And the reason I like it is ultimately diagnosis. The process of diagnosis is pragmatic, it's to, you know, to do good things, hopefully, for our patients.

So kidney disease back in the middle of the 19th century would've been called renal dropsy. Dropsy is actually a very ancient word just meaning getting fluid on, or keeping fluid on, blowing up like a balloon. And, by the middle of the 19th century, it was understood that sometimes the kidneys failed, causing this backup of fluid and there were treatments for it. They were not particularly pretty treatments. They would do these things, using things called Southey tubes, basically giant trocars or tubes that they would insert into the legs and fluid would flow out of it. They might use arsenic-based diuretics to get some fluid off all very, very nasty treatments in this same period. They started, well, they were already doing autopsies, but they started doing kidney biopsies on living patients. And physicians, Richard Bright is the famous one, started to note that there were actually pathological changes in patients who had kidney disease that showed how developed the disease was.

And they would perform these kidney biopsies to tell a patient, to give them important prognostic information about how much time they might have until their kidneys finally failed. Because when the kidneys failed that's the end, there's no effective treatment. This was called Bright's disease and Bright's disease persisted until really the almost middle of the 20th century. Well, in the middle of the 20th century, we started to think differently about kidney disease. And we started to think about from a more functional status, like how well are the kidneys working? What is the glomerular filtration rate, GFR? Initially it's an academic discussion, but then all of a sudden hemodialysis comes along, right? We have machines now that can serve as external kidneys. And the GFR becomes a proxy for how we determine when somebody gets hemodialysis. And then later on, we would come up with new therapies that could slow the progression of kidney disease.

Well, now all of a sudden Bright's disease is no longer helpful. We don't care about what pathological changes are happening. We care about how effective the kidneys are working. Why? Because there's a treatment that we can use to actually treat end stage renal disease. And then later on in the eighties on there's medications that we can give to slow the progression of renal disease. So even though renal dropsy, which is based on physical exam findings, Bright's disease, which is based on pathological findings and CKD, which is based on, you know, chemical findings, even though they're all based on real things, you can see how the conception of this disease has changed over 150 years to pragmatically update because of the things that we're able to now do for our patients.

[SA] Bedside rounds is going into, I think it's sixth episode about diagnosis. We already have sort of an idea, but what about the idea of diagnosis is so fascinating to you?

[AR] Oh yeah. So this is great. So I love medical history. I love, I think obviously there are many great medical historians out there who think about this. So I don't wanna, I'm definitely not alone in this, but in medicine, in practical medicine, when we think about medical history, often it's talked about in terms of advancement, right? Oh, we understood this disease better. We develop new therapies and that's incredibly important. What I like about diagnosis is that it's not a thing. It's not an object. It is a construct. It is a metacognitive construct and it interacts with the science, right? It interacts with our scientific understanding of the human body. It interacts with how we classify diseases, but it refers to invisible things that happen in our heads. And increasingly in the heads of the tools that we've built, right?

The computers, the decision tools that we've built, maybe eventually artificial intelligence that we might use. And this thought like we have a very, I think, naive understanding. And when I say we I'm referring to just the practice of medicine in general, that, oh, it's just pattern recognition. You just look at all the facts and you fit it into one of these disease categories. But when you actually start to dig, even just superficially deep into a diagnosis, you can see the way that we've conceived the process has changed dramatically over time. And it continues to change in ways that are I think exciting and in ways that are invisible to us, like metacognition by its very nature it's meta, right? We don't think about it. It's, it's just something that we do. And, and to me, I just, I love that.

[SA] So I wonder given the fact that you've made very clear that all of the sands are shifting all the time, how do you communicate this to your medical students and how are they supposed to process that information given that they're overwhelmed with even just learning what is in vogue right now?

[AR] Oh my goodness. You asked a really tough question. That's a great question. Right, so the conception here is when we teach medical students about diagnosis, we're really clear cut. Right? We have very clear categories. When we teach them about test characteristics, we teach them, what would we call them? Comforting half-truths, right. We teach them that test characteristics are immutable. You know, you just, I, so for the audience, there are apps that you can download where you're like, oh, they have an elevated JVP. Yes. They have, uh, lower extremedema. Yes. And it shoots out the probability of heart failure and it makes diagnosis seem incredibly scientific, right? Like if you just collect enough data, you'll be able to put it all in and spit out a diagnosis. Now the reality, that's not true at all. So how do I express this to my students?

There's no easy way, right? Because you don't want people to despair. Because the truth is, is it's not despairing, but it's a lot more complicated than we initially presented. What I try to do is practice what I preach and think out loud and talk about when we get a farin (?), why, you know, the farin of 200, I'm pretty sure is, you know, anemia of chronic inflammation, but this 40, maybe we don't know. And I try to demonstrate what my thinking is because ultimately I think that's the only way we can do this. I mean, we cannot, I know every medical student hates two-by-two tables. Everyone just forces and memorizes them to, to get past Step One. And now the Step One's pass/fail. I imagine to get past Step Two, but it actually is really important.

And I don't want people to feel so cynical that they're like, I'm not even going to learn that. So I try to model, I walk the walk. I do the same thing with patients too. I talk about uncertainty all the time with patients, uncertainty in my diagnostics. I talk about empiric therapy with pneumonia. A patient with pneumonia is gonna get a lecture on epistemology, because really for me, I explain to them my decision doesn't come down to you have a pneumonia. Yes, definitely. You have a pneumonia. No, it comes down to: Is my probability of you having a pneumonia worth the risk of exposing you to antibiotics? And I try to do that in lots of the things I do, which is why I spend so much time talking to people.

[SA] Thank you so much, Adam. This has been fascinating.

[AR] Oh, thank you so much for having me, Sarah. This is awesome.

[SA] Listeners, thank you for tuning in to the Proto podcast. Today's podcast was produced by Joshua Krisch, Bradley Klein and Jason Anthony. Thanks also to our technical directors, Adam Keller and Nathan

Marcus. Subscribe to the Proto podcast on iTunes and Stitcher and follow us on Facebook, Twitter, and Instagram. Stay safe. See you next time.