DAVID SONTAG: Today we'll be talking about risk stratification. After giving you a broad overview of what I mean by risk stratification, we'll give you a case study which you read about in your readings for today's lecture coming from early detection of type 2 diabetes. And I won't be, of course, repeating the same material you read about it in your readings. Rather I'll be giving some interesting color around what are some of the questions that we need to be thinking about as machine learning people when we try to apply machine learning to problems like this. Then I'll talk about some of the subtleties. What can go wrong with machine learning based approaches to risk stratification? And finally, the last half of today's lecture is going to be a discussion.

So about 3:00 PM, you'll see a man walk through the door. His name is Leonard D'Avolio. He is a professor at Brigham Women's Hospital. He also has a startup company called Sift, which is working on applying risk stratification now, and they have lots of clients. So they've been really deep in the details of how to make this stuff work. And so we'll have an interview between myself and him, and we'll have opportunity for all of you to ask questions as well. And that's what I hope will be the most exciting part of today's lecture.

Then going on beyond today's lecture, we're now in the beginning of a sequence of three lectures on very similar topics. So next Thursday, we'll be talking about survival modeling. And you can think about it as an extension of today's lecture, talking about what you should do if your data has centering, which I'll define for you shortly. Although today's lecture is going to be a little bit more high level, next Thursday's lecture is where we're going to really start to get into mathematical details about how one should tackle machine learning problems with centered data. And then the following lecture after that is going to be on physiological data, and that lecture will also be much more technical in nature compared to the first couple of weeks of the course.

So what is risk stratification? At a high level, you think about risk stratification as a way of taking in the patient population and separating out all of your patients into one of two or more categories. Patients with high risk, patients with low risk, and maybe patients somewhere in the middle.

Now the reason why we might want to do risk stratification is because we usually want to try to act on those predictions. So the goals are often one of coupling those predictions with known

interventions. So for example, patients in the high risk pool-- we will attempt to do something for those patients to prevent whatever that outcome is of interest from occurring.

Now risk stratification is quite different from diagnosis. Diagnosis often has very, very stringent criteria on performance. If you do a mis-diagnosis of something, that can have very severe consequences in terms of patients being treated for conditions that they didn't need to be treated for, and patients dying because they were not diagnosed in time.

Risk stratification you think of as a little bit more fuzzy in nature. We want to do our best job of trying to push patients into each of these categories-- high risk, low risk, and so on. And as I'll show you throughout today's lecture, the performance characteristics that we'll often care about are going to be a bit different. We're going to look a bit more at quantities such as positive predictive value. Of the patients we say are high risk, what fraction of them are actually high risk? And in that way, it differs a bit from diagnosis.

Also as a result of the goals being different, the data that's used is often very different. In risk stratification, often we use data which is very diverse. So you might bring in multiple views of a patient. You might use auxiliary data such as patients' demographics, maybe even socioeconomic information about a patient, all of which very much affect their risk profiles but may not be used for a unbiased diagnosis of the patient.

And finally in today's economic environment, risk stratification is very much targeted towards reducing cost of the US health care setting. And so I'll give you a few examples of risk stratification, some of which have cost as a major goal others which don't. The first example is that of predicting an infant's risk of severe morbidity. So this is a premature baby. My niece, for example, was born three months premature. It was really scary for my sister and my whole family. And the outcomes of patients who are born premature have really changed dramatically over the last century. And now patients who are born three months premature, like my niece, actually can survive and do really well in terms of long term outcomes.

But of the many different inventions that led to these improved outcomes, one of them was having a very good understanding of how risky a particular infant might be. So a very common score that's used to try to characterize risk for infant birth, generally speaking, is known as the Apgar score. For example when my son was born, I was really excited when a few seconds after my son was delivered, the nurse took out a piece of paper and computed the Apgar score. I studied that, really interesting, right? And then I got back to some other things that I

had to do.

But that score isn't actually as accurate as it could be. And there is this paper, which we'll talk about in a week and a half, by Suchi Saria who's a professor at Johns Hopkins, which looked at how one could use a machine learning based approach to really improve our ability to predict morbidity in infants.

Another example, which I'm pulling from the readings for today's lecture, has to do with-- for patients who come into the emergency department with a heart related condition, try to understand do they need to be admitted to the coronary care unit? Or is it safe enough to let that patient go home and be managed by their primary care physician or their cardiologist outside of the hospital setting?

Now that paper, you might have all noticed, was from 1984. So this isn't a new concept. Moreover, if you look at the amount of data that they used in that study, it was over 2,000 patients. They had a nontrivial number of variables, 50 something variables. And they used a non-trivial machine learning algorithm. They used logistic regression with a feature selection built in to prevent themselves from over fitting to the data. And the goal there was very much cost oriented. So the premise was that if one could quickly decide these patients who've just come to the ER are not high risk and we could send them home, then we'll be able to reduce the large amount of cost associated with those admissions to coronary care units.

And the final example I'll give right now is that of predicting likelihood of hospital readmission. So this is something which is getting a real lot of attention in the United States health care space over the last few years because of penalties which the US government has imposed on hospitals who have a large number of patients who have been released from the hospital, and then within the next 30 days readmitted to the hospital. And that's part of the transition to value based care, which Pete mentioned in earlier lectures.

And so the premise is that there are many patients who are hospitalized but are not managed appropriately on discharge or after discharge. For example, maybe this patient who has a heart condition wasn't really clear on what they should have done when they go home. For example, what medications should they be taking? When should they follow up with their cardiologist? What things they should be looking out for, in terms of warning signs that they should go back to the hospital or call their doctor for. And as a result of that poor communication, it's conjectured that these poor outcomes might occur.

So if we could figure out which of the patients are likely to have those readmissions, and if we could predict that while the patients are still in the hospital, then we could change the way that discharge is done. For example, we could send a nurse or a social worker to talk to the patient. Go really slowly through the discharge instructions. Maybe after the patient is discharged, one could have a nurse follow up at the patient's home over the next few weeks. And in this way, hopefully reduce the likelihood of that readmission.

So at a high level, there's the old versus the new. And this is going to be really a discussion throughout the rest of today's lecture. What's changed since that 1984 article which you read for today's readings? Well, the traditional approaches to risk stratification are based on scoring systems. So I mentioned to you a few minutes ago, the Apgar scoring system is shown here.

You're going to say for each of these different correct criteria-- activity, pulse, grimace, appearance, respiration-- you look at the baby, and you say well, activity is absent. Or maybe they're active movement. Appearance might be pale or blue, which would get 0 points, or completely pink which gets 2 points. And for each one of these answers, you add up the corresponding points. You get a total number of points. And you look over here and you say, OK, well if you have a 0 to 3 points, the baby is at severe risk. If they have 7 to 10 points, then the baby is low risk.

And there are hundreds of such scoring rules which have been very carefully derived through studies not dissimilar to the one that you read for today's readings, and which are actually widely used in the health care system today. But the times have been changing quite rapidly in the last 5 to 10 years. And now, what most of the industry is moving towards are machine learning based methods that can work with a much higher dimensional set of features and solve a number of key challenges of these early approaches.

First-- and this is perhaps the most important aspect, they can fit more easily into clinical workflows. So the scores I showed you earlier are often done manually. So one has to think to do the score. One has to figure out what the corresponding inputs are. And as a result of that, often they're not used as frequently as they should be. Second, the new machine learning approaches can get higher accuracy potentially, due to their ability to use many more features than the traditional pitches. And finally, they can be much quicker to drive.

So all of the traditional scoring systems had a very long research and development process that led to their adoption. First, you gather the data. Then you build the models. Then you

check the models. Then you do an evaluation in one hospital. Then you do a prospective evaluation in many hospitals. And each one of those steps takes a lot of time.

Now with these machine learning based approaches, it raises the possibility of a research assistant sitting in a hospital, or in a computer science department, saying oh, I think it would be really useful to derive a score for this problem. You take data that's available. You apply your machine learning algorithm. And even if it's a condition or an outcome which occurs very infrequently, if you have access to a large enough data set you'll be able to get enough samples in order to actually predict that somewhat very narrow outcome. And so as a result, it really opens the door to rethinking about the way that risk stratification can be used.

But as a result, there are also new dangers that are introduced. And we'll talk about some of those in today's lecture, and we'll continue to talk about those in next Thursday's lecture. So these models are being widely commercialized. Here is just an example from one of many companies that are building risk stratification tools. This is from Optum. And what I'm showing you here is the output from one of their models which is predicting COPD related hospitalizations. And so you'll see that this is a population level view. So for all of the patients who are of interest to that hospital, they will score the patient-- using either one of the scores I showed you earlier, the manual ones, or maybe a machine learning based model-- and they'll be put into one of these different categories depending on the risk level.

And then one can dig in deeper. So for example, you could click on one of those buckets and try to see well, who are the patients that are highest at risk. And what are some potentially impactible aspects of those patients' health? Here, I'm showing you for a slightly different problem that are predicting high risk diabetes patients. And you see that for each patient, we're listing the number of A1C tests, the value of the last A1C test, the day that it was performed. And in this way, you could notice oh, this patient is at high risk of having diabetes. But look, they haven't been tracking their A1C. Maybe they have uncontrolled diabetes. Maybe we need to get them into the clinic, get their blood tested, see whether maybe they need a change in medication, and so on. So in this way, we can stratify the patient population and think about interventions that can be done for that subset of them.

So I'll move now into a case study of early detection of type 2 diabetes. The reason why this problem is of importance is because it's estimated that there are 25% of patients with undiagnosed type 2 diabetes in the United States. And that number is equally large as you go to many other countries internationally. So if we can find patients who currently have diabetes

or are likely to develop diabetes in the future, then we could attempt to impact them.

So for example, we could develop new interventions that can prevent those patients from worsening in their diabetes progression. For example, weight loss programs or getting patients on first line diabetic treatments like Metformin. But the key problem which I'll be talking about today is really, how do you find that at risk population?

So the traditional approach to doing that is very similar to that Apgar score. This is a scoring system used in Finland which asks a series of questions and has points associated with each answer. So what's the age of the patient? What's their body mass index? Do they eat vegetables, fruit? Have they ever taken anti hypertension medication? And so on, and you get a final score out, right? Lower than 7 would be 1 in 100 risk of developing type 2 diabetes. Higher than 20 is very high risk. 1 in 2 people will develop type 2 diabetes in the next 10 years.

But as I mentioned, these scores haven't had the impact that we had hoped that they might have. And the reason really is because they haven't been actually used nearly as much as they should be. So what we will be thinking through is, can we change the way in which risk stratification is done? Rather than it having to be something which is manually done, when you think to do it, we can make it now population wide.

We could, for example, take data that's already available from a health insurance company, use machine learning. Maybe we don't have access to all of those features I showed you earlier. Maybe we don't know the patient's weight, but we will use machine learning on the data that we do have to try to find other surrogates of those things we don't have, which might predict diabetes risk. And then we can apply it automatically behind the scenes for millions of different patients and find the high risk population and perform interventions for those patients. And by the way, the work that I'm telling you about today is work that really came out of my lab's research in the last few years.

So this is an example going back to the set of stakeholders, which we talked about in the first lecture. This is an example of a risk stratification being done at the payer level. So the data which is going to be used for this problem is administrative data, data that you typically find in health insurance companies. So I'm showing you here a single patient's timeline and the type of data that you would expect to be available for that patient across time.

In red, it's showing their eligibility records. When had they been enrolled in that health insurance? And that's really important, because if they're not enrolled in the health insurance

on some month, then the lack of data for that patient isn't because nothing happened. It's because we just don't have visibility into it. It's missing. In green, I'm showing medical claims which are associated with diagnosis codes that Pete talked about last week, procedure codes, CPT codes. We know what the specialist was that the patient went to see, like cardiologists, primary care physician, and so on. We know where the service was performed, and we know when it was performed.

And then from pharmacy, we have access to medication records shown in the top right there. We know what medication was prescribed, and we have it coded to the NDC code-- National Drug Code, which Pete talked about again last Tuesday. We know the number of days' supply of the medication, the number of refills that are available still, and so on.

And finally, we have access to laboratory tests. Now traditionally, health insurance companies only know what tests were performed because they have to pay for that test to be performed. But more and more, health insurance companies are forming partnerships with companies like Quest and LabCorps to actually get access also to the results of those lab tests. And in the data set that I'll tell you about today, we actually do have those lab test results as well.

So what are these elements for this population? This population comes from Philadelphia. So if we look at the top diagnosis codes, for example, we'll see that of 135,000 patients who had laboratory data, there were over 400,000 different diagnosis codes for hypertension. You'll notice that's greater than the number of people. That's because they occurred multiple times across time. Other common diagnosis codes included hyperlipidemia, hypertension, type 2 diabetes. And you'll notice that there's actually quite a bit of interesting detail here. Even in diagnosis codes, you'll find things that sound more like symptoms-- like fatigue, which is over here. Or you also have records of procedures, in many cases. Like they got a vaccination for influenza.

Here's another example. This is now just telling you something about the broad statistics of laboratory tests in this population. Creatinine, potassium, glucose, liver enzymes are all the most popular lab tests. And that's not surprising, because often there is a panel called the CBC panel which is what you would get in your annual physical. And that has many of these top laboratory test results. But then as you look down into the tail, there are many other laboratory test results that are more specialized in nature. For example, hemoglobin A1C is used to track roughly 3 month average of blood glucose and is used to understand a patient's diabetes status.

So that's just to give you a sense of what is the data behind the scenes. Now let's think, how do we really derive-- how do we tackle-- how do we formulate this risk stratification problem as a machine learning problem? Well today, I'll give you one example of how to formulate it as a machine learning problem. But in Tuesday's lecture, I'll tell you several other ways.

Here, we're going to think about a reduction to binary classification. We're going to go back in time. We're going to pretend it's January 1, 2009. We're going to say suppose that we had run this risk stratification algorithm on every single patient on January 1, 2009. We're going to construct features from the data in the past, so the past few years. We're going to predict something about the future. And there many things you could attempt to predict about the future. I'm showing you here 3 different prediction tasks corresponding to different gaps-- a 0 year gap, a 1 year gap, and a 2 year gap. And for each one of these, it asks will the patient newly develop type 2 diabetes in that prediction window?

So for example, for this prediction task we're going to exclude patients who have developed type 2 diabetes between 2009 and 2011. And we're only going to count as positives patients who get newly diagnosed with type 2 diabetes between 2011 and 2013. And one of the reasons why you might want to include a gap in the model is because often, there's label leakage. So if you look at the very top set up, often what happens is a clinician might have a really good idea that the patient might be diabetic, but it's not yet coded in a way which our algorithms can pick up.

And so on January 1, 2009 the primary care physician for the patient might be well aware that this patient is diabetic, might already be doing interventions based on it. But our algorithm doesn't know that, and so that patient, because of the signals that are present in the data, is going to at the very top of our prediction list. We're going to say this patient is someone you should be going after. But that's really not an interesting patient to be going after, because the clinicians are probably already doing interventions that are relevant for that patient. Rather, we want to find the patients where the diabetes might be more unexpected.

And so this is one of the subtleties that really arises when you try to use retrospective clinical data to derive your labels to use within machine learning for risk stratification. So in the result I'll tell you about, I'm going to use a 1 year gap. Another problem is that the data is highly censored. So what I mean by censoring is that we often don't have full visibility into the data for a patient. For example, patients might have only come into the health insurance in 2013,

and so January 1, 2009 we have no data on them. They didn't even exist in the system at all.

So there are two types of censoring. One type of censoring is called left censoring. It means when we don't have data to the left, for example in the feature construction window. Another type of censoring is called right censoring. It means when we don't have data about the patient to the right of that time line. And for each one of these in our work here, we tackle it in a different way. For left centering, we're going to deal with it. We're going to say OK, we might have limited data on patients. But we will use whatever data is available from the past 2 years in order to make our predictions. And for patients who have less data available, that's fine. We have sort of a more sparse feature vector.

For right centering, it's a little bit more challenging to deal with in this binary reduction, because if you don't know what the label is, it's really hard to use within, for example, a supervised machine learning approach. In Tuesday's lecture, I'll talk about a way to deal with right censoring. In today's lecture, we're going to just ignore it. And the way that we'll ignore it is by changing the inclusion and exclusion criteria. We will exclude patients for whom we don't know the label.

And to be clear, that could be really problematic. So for example, imagine if you go back to this picture here. Imagine that we're in this scenario. And imagine that if we only have data on a patient up to 2011, we remove them from the data set, OK? Because we don't have full visibility into the 2010 to 2012 time window.

Well, suppose that exactly the day before the patient was going to be removed from the data set-- right before the data disappears for the patient because, for example, they might change health insurers-- they were diagnosed with type 2 diabetes. And maybe the reason why they changed health insurers had to do with them being diagnosed with type 2 diabetes. Then we've excluded that patient from the population, and we might be really biasing the results of the model, by now taking away a whole set of the population where this model would've been really important to apply. So thinking about how you really do this inclusion exclusion and how that changes the generalizability of the model you get is something that should be at the top of your mind.

So the machine learning algorithm used in that paper which you've read is L1 regularized logistic regression. One of the reasons for using L1 regularized logistic regression is because it provides a way to use a high dimensional feature set. But at the same time, it allows one to

do feature selection. So I'll go more into detail on that in just a moment. All of you should be familiar with the idea of formulating machine learning as an optimization problem where you have some loss function, and you have some regularization term-- w, in this case, as the weights of your linear model, which we're trying to learn. For those of you who've seen support vector machines before, support vector machines will use what's called L2 regularization where we'll be putting a penalty on the L2 norm of the weight vector.

Instead, what we did in this paper is used L1 regularization. So this penalty is defined over here. It's summing over the features and looking at the absolute value for each of the weights and summing those up. So one of the reasons why L1 regularization has what's known as a sparsity benefit can be explained by this picture. So this is just a demonstration by sketch. Suppose that we're trying to solve this optimization problem here.

So this is the level set of your loss function. It's a quadratic function. And suppose that instead of adding on your regularization as a second term to your optimization problem, you were to instead put in a constraint. So you might say we're going to minimize the loss subject to the L1 norm of your weight vector being less than 3. Well, then what I'm showing you here is weight space. I'm showing you 2 dimensions. This x-axis is weight 1. This y-axis is weight 2.

And if you put an L1 constraint-- for example, you said that the sum of the absolute values of weight 1 and weight 2 have to be equal to 1-- then the solution space has to be along this diamond. On the other hand, if you put an L2 constraint on your weight vector, then it would correspond to this feasibility space. For example, this would say something like the L2 norm over the weight vector has to be equal to 1. So it would be a ball, saying that the radius has to always be equal to 1.

So suppose now you're trying to minimize that objective function, subject to the solution having to be either on the ball, which is what you would do if you were optimizing the L2 norm, versus living on this diamond, which is what would happen if you're optimizing the L1 norm. Well, the optimal solution is going to be in essence the closest point along the circle, which gets as close as possible to the middle of that level set. So over here, the closest point is that 1. And you'll see that this point has a non-zero w1 and w2. Over here, the closest point is over here. Notice that has a zero value of w1 and a non-zero value of w2, thus it's found a sparser solution than this one. So this is just to give you some intuition about why using L1 regularization results in sparse solutions to your optimization problem.

And that could be beneficial for two purposes. First, it can help prevent over fitting in settings where there exists a very good risk model that uses a small number of features. And to point out, that's not a crazy idea that there might exist a risk model that uses a small number of features, right? Remember, think back to that Apgar score or the FINDRISC, which was used to predict diabetes in Finland. Each of those had only 5 to 20 questions. And based on the answers to those 5 to 20 questions, one could get a pretty good idea of what the risk is of that patient, right? So the fact that there might be a small number of features that are together sufficient is actually a very reasonable prior. And it's one reason why L1 regularization is actually very well suited to these types of risk stratification problems on this type of data.

The second reason is one of interpretability. If one wants to then ask, well, what are the features that actually were used by this model to make predictions? When you find only 20 or a few features, you can enumerate all of them and look to see what they are. And in that way, understand what is going on into the predictions that are made. And that also has a very big impact when it comes to translation.

So suppose you built a model using data from this health insurance company. And this health insurance company just happened to have access to a huge number of features. But now you want to go somewhere else and apply the same model. If what you've learned is a model with only a few hundred features, you're able to dwindle it down. Then it provides an opportunity to deploy your model much more easily. The next place you go to, you only need to get access to those features in order to make your predictions.

So I'll finish up in the next 5 minutes in order to get to our discussion with Leonard. But I just want to recap what are the features that go into this model, and what are some of the valuations that we use. So the features that we used here were ones that were designed to take into consideration that there is a lot of missing data for patients. So rather than think through do we impute this feature, do we not impute this feature, we simply look to see were these features ever observed? So we choose our feature space in order to already account for the fact that there's a lot missing.

For example, we look to see what types of specialists has this doctor seen in the past, been to in the past? For every possible specialist, we put a 1 in the corresponding dimension if the patient has seen that type of specialist and 0 otherwise. For the top 1,000 most common medications, we look to see has the patient ever taken his medication, yes or no? And again, 0 or 1 in the corresponding dimension.

For laboratory tests, that's where we do something which is a little bit different. We look to see, first of all, was a laboratory test ever administered? And then we say OK, if it was administered, was the result ever low, out of bounds on the lower side? Was the result ever high? Was the result ever normal? Is the value increasing? Is the value decreasing? Is the value fluctuating? I noticed that each one of these quantities is well-defined, even for patients who don't ever have any laboratory test results available, right? The answer would be 0, it was never administered. And 0, it was never low. 0, it was never high, and so on. OK?

AUDIENCE: Is the value increasing? Is it every time, or how do you define?

DAVID SONTAG: So increasing here-- first of all, if there is only a single value observed then it's 0. If there were at least 2 values observed, then you look to see was there ever any adjacent pair of observations where the second one was higher than the first one? That's the way it was defined here.

AUDIENCE: Then it has increased and then decreased. You put 1 and 1 on the [INAUDIBLE].

DAVID SONTAG: Correct. That's what we did here. And it's extremely simple, right? So there are lots of better ways that you could do this. And in fact, this is an example which we'll come back to perhaps a little bit in the next lecture and then more in subsequent lectures when we talk about using recurrent neural networks to try to summarize time series data. Because one could imagine that using such an approach could actually automatically learn such features.

AUDIENCE: Just to double check, is fluctuating one of the other two [INAUDIBLE]?

DAVID SONTAG: Fluctuating is exactly the scenario that was just described. It can go up, and then it goes down. Has to do both, yeah. Yep?

AUDIENCE: It said in the first question, [INAUDIBLE] together. Was the test ever administered [INAUDIBLE]? And the value you have there is 1.

DAVID SONTAG: Correct. So indeed, there is a huge amount of correlation between these features. If any of these were 1, then this is also going to be 1.

AUDIENCE: Especially the results.

DAVID SONTAG: Yeah, but you would still want to include this 1 in here. So imagine that all of these were 0. You don't know if they're 0 because these things didn't happen or because the test was never

performed.

AUDIENCE: Are the low, high, normal--

DAVID SONTAG: They're just binary indicators here, right?

AUDIENCE: Doesn't it have to fit into one category?

DAVID SONTAG: Well, no. Oh, I see what you're saying. So you're saying if the result was ever present, then it would be at least 1 of these 3. Maybe. It gets into some of the technical details which I don't remember right now. It was a good question.

And this is the next most really important detail. The way I just described this, there was no notion of time in that. But of course when these things happened can be really important. So the next thing we do is we re-compute all of these features for different time buckets. So we compute them for the last 6 months of history, for the last 24 months of history, and then for all of the past history. And we can catenate together all of those feature vectors and what you get out. In this case, it was something like a 42,000 dimensional feature vector. By the way, it's 42,000 dimensional and not higher because the features that we used for diagnosis codes for this paper were not temporal in nature. And one could easily make them temporal in nature, in which case it'd be more like 60,000 features.

I'm going to skip over the deriving labels and get back to that next time. I just want to briefly talk about how does one evaluate these types of models. And I'll give you one view on evaluations, and shortly we'll hear a very different type of view. So here, what I'm showing you are the variables that have been selected by the model and have non-zero weight. So for example, the very top you see impaired fasting glucose, which is used by the model. It's not surprising because we're trying to predict is the patient likely to develop type 2 diabetes.

Now you might ask, if a patient has a diagnosis code for impaired fasting glucose aren't they already diabetic? Shouldn't they have been excluded? And the answer is no, because there are also patients who are pre-diabetic in this data set, who have been intentionally included because we don't know which of them are going to go on to develop type 2 diabetes. And so this is an indicator that the patient has been previously flagged as being pre-diabetic. And it obviously makes sense that would be at the very top of the predictive variables.

But there are also many things that are a little bit less obvious. For example, here we see

obstructive sleep apnea and esophageal reflux as being chosen by the model to be predictive of the patient developing type 2 diabetes. What we would conjecture is that those variables, in fact, act as surrogates for the patient being obese. Obesity is very seldom coded in commercial health insurance claims. And so with this variable, despite the fact that the patient might be obese, if this variable is not observed then patients who are obese often have what's called sleep apnea. So they might stop breathing for short periods of time during their sleep. And so that then would be a sign of obesity.

So I talked about how the criteria which we use to evaluate risk stratification models are a little bit different from the criteria used to evaluate diagnosis models. Here I'll tell you one of the measures that we often use, and it's called positive predictive value. So what we'll do is look at after you've learned your model. Look at the top 100 predictions, top 1,000 predictions, top 10,000 predictions, and look to see what fraction of those patients went on to actually develop type 2 diabetes. Now of course, this is done using held up data. Now the reason why you might be interested in different levels is because you might want to target different interventions depending on the risk and cost.

For example, a very low cost intervention-- one of the ones that we did-- was sending a text message to patients who are suspected to have high risk of developing type 2 diabetes. If they've not been to see their eye doctor in the last year, we send them a text message saying maybe you want to go see your eye doctor. Remember, you get a free eye checkup. And this is a very cheap intervention, and it's a very subtle intervention. The reason why it can be effective is because patients who develop type 2 diabetes, once that diabetes progresses it leads to something called diabetic retinopathy, which is often caught in an eye exam. And so that could be one mechanism for patients to be diagnosed.

And so since it's so cheap, you could do it for 10,000 people. So you take the 10,000 most risky people. You apply the intervention for them, and you look to see which of those people actually had developed diabetes in the future. In the model that I showed you, 10% of that population went on to develop type 2 diabetes 1 to 3 years from then. The comparison point I'm showing you here, this blue bar, is if you used a model which is derived using a very small number of features, so not a machine learning based approach. And there, only 6% of the people went on to develop type 2 diabetes from the top 10,000.

On the other hand, other interventions you might want to do are much more expensive. So for example, you might only be able to do that intervention for 100 people because it costs so

much money, and you have a limited budget as a health insurer. And so for those people, you could ask well, what is the positive predictive value of those top 100 predictions? And here, that was 15% using the machine learning based model and less than half of that using the more traditional approach.

So I'm going to stop here. There's a lot more that I can and will say. But I'll have to get to it in next Thursday's lecture, because I'd like our guest to come down, and we will have a bit of a discussion. To be clear, this is the first time that we've ever had this type of class interaction which is why, by the way, I ran a little bit late. I hadn't ever done something like this before. So it's an experiment. Let's see what happens. So, do you say Leonard?

LEONARD D'AVOLIO: Len's fine.

DAVID SONTAG: Len, OK. So Len, could you please introduce yourself?

LEONARD D'AVOLIO: Sure. My name is Len D'Avolio. I'm an assistant professor at Harvard Medical School. I am also the CEO and founder of a company called Sift. Do you want a little bit of background or no?

DAVID SONTAG: Yeah, a little bit of background.

LEONARD D'AVOLIO: Yeah, so I've spent probably the last 15 years or so trying to help health care learn from its data in new ways. And of all the fields that need your help, I would say health care for both societal, but also just from a where we're at with our ability to use data standpoint is a great place for you guys to invest your time. I've been doing this for government, in academia as a researcher, publishing papers. I've been doing this for non-profits in this country and a few others.

But every single project that I've been a part of has been an effort to bring in data that has always been there, but we haven't been able to learn from until now. And whether that's at the VA building out there, genomic science infrastructure, recruiting and enrolling a million veterans to donate their blood and their EMR, or at Ariadne Labs over out of Harvard School of Public Health and the Brigham, improving childbirth in India-- it's all about how can we get a little bit better over and over again to make health care a better place for folks.

DAVID SONTAG: So tell me, what is risk stratification from your perspective? Defining that I found to be one of the most difficult parts of today's lecture.

LEONARD D'AVOLIO: Well, thank you for challenging me with it.

[LAUGHTER]

So it's a rather generic term, and I think it depends entirely on the problem you're trying to solve. And every time I go at this, you really have to ground yourself in the problem that you're trying to solve. Risk could be running out of a medical supply in an operating room. Risk could be an Apgar score. Risk could be from pre-diabetic to diabetic. Risk could be an older person falling down in their home.

So really, what is it to me? I'm very much caught up in the tools analogy. These are wonderful tools with which a skilled craftsman surrounded by others that have skills could go ahead and solve very specific problems. This is a hammer. It's one that we spend a lot of time refining and applying to solve problems in health care.

DAVID SONTAG: So why don't you tell us about some of the areas where your company has been applying risk stratification today at a very high level. And then we'll choose on of them to dive a bit deeper into.

LEONARD D'AVOLIO: Sure. So the way we describe what we do is it's performance improvement. And I'm just giving you a little background, because it'll tell you which problems I'm focused on. So it's performance improvement, and to be candid, the types of things we like to improve the performance of are how do we keep people out of the hospital. I'm not going to soapbox on this too much, but I think it matters.

Like the example that you gave that you were employed to help solve was by an insurer, and insurance companies-- there's probably 30 industries in health care. It's not one industry. And every one of them has different and oftentimes competing incentives. And so the most logical application for these technologies is to help do preventative things. But only about, depending on your math, between 8% and 12% of health care is financially incentivized to do preventative things. The rest are the hospitals and the clinics. And when you think of health care, you probably think of those types of organizations. They don't typically pay to keep you out of those facilities.

DAVID SONTAG: So as a company, you know, you've got to make a profit of entry. So you need to focus on the

ones where there's a financial incentive.

LEONARD D'AVOLIO: You focus on where there's a financial incentive. And in my case, I wanted to build a company where the financial incentive aligned with keeping people healthy.

DAVID SONTAG: So what are some of these examples?

LEONARD D'AVOLIO: Sure. So we do a lot with older populations. With older populations, it becomes very important to understand who care managers should approach, because their risk levels are rising. A lot of risk stratification, the old way that you described, identifies people that are already at their most acute. So it's sort of skating to where the puck has been. You're getting attention because you are at the absolute peak of your acuity.

We're trying to help care management organizations find people that are rising risk. And even when we do that, we try to get-- I mean, the power of these technologies is to move away from one size fits all. So when we think about rising risk, we think about in a behavioral health environment, it is the rising risk of an inpatient psychiatric admission. That is a very specific application. There are things we can do about it. As opposed to risk, which if you think about what's being done in other industries, Amazon does not consider us all consumers. There are individuals that are very likely to react to certain offers at certain times.

And so we're trying to bring this sort of more granular approach into health care, where we sit with teams and they're used to just having generic risk scores. We're trying to help them think through which older people are likely to fall down. We do work in diabetes also, so which children with type 1 diabetes shouldn't just be scheduled for an appointment every 3 months, but you should go to them right now?

So those are some examples, but the themes are very consistent. It's helping organizations move away from rather generic, one size fits all toward what are the more actionable. So even graduation from care management, because now you should be having serious illness conversations because you're nearing end of life, or palliative care referrals, or hospice referrals.

DAVID SONTAG: OK, so I want to choose a single one to dive into. And I want to choose one that you've worked on the longest and where you're already doing at least the initial parts of an evaluation of it. And so I think when we talked on the phone, psyche ER was one of those examples. Tell us a bit about that one.

LEONARD D'AVOLIO: Yeah. Well, I'll just walk you through the problem to be solved.

DAVID SONTAG: Please, yeah.

LEONARD D'AVOLIO: Sure. So we work with a large behavioral health care organization. They are contracted by health plans, in effect, to treat people that have mental health challenges. And the traditional way of identifying anyone for care management is again, you get a risk score. When you sort the highest ranking in terms of odds ratio variables, it's because you were already admitted, because you're older, because you have more medications. So they were using a similar approach, finding the most acute people.

So the very first thing we do in all of our engagements is an understanding. Where is the greatest opportunity? And this has very little to do with machine learning. It's just what's happening today? Where are these things happening? Who is caring for these folks? Everyone wants to reduce hospital admissions. But there's a difference between hospital admissions because you're not taking your meds, and hospital admissions because you're addicted to opioids, and hospital admissions because you have chronic complex bipolar schizophrenia.

So we wanted to first understand well, where is the greatest cost? What types of things are happening most frequently? And then you want to have the clinical team tell you well, these are the types of resources we have. We have people that can address these issues, or we have interventions designed to solve these problems. And so you bring together where is the greatest possible return on your investment from both a data standpoint, a financial standpoint, but also and we can do something about it. After you do that, it's only then-- after you have full agreement from executive teams-- that this is the very narrow thing that we think we can address. Then we begin to apply machine learning to try to solve the problem.

DAVID SONTAG: So what did that funnel lead to? What did you decide was the thing to address?

LEONARD D'AVOLIO: Yeah, it was tried to reduce inpatient psychiatric admissions. And even then, the traditional way of reducing admissions-- just because it came out of this tradition of 30 day readmissions-- has always been thought of in terms of 30 days out. But when we interviewed the teams, they said actually for this particular condition it takes us more like 90 days to be able to have an impact. And so that clinical understanding mixed with what we have the resources to

address, that's what steers then the application of machine learning to solve a specific problem.

DAVID SONTAG: OK, so psychiatric inpatient admission-- so these are patients who come to the ER for some psychiatric related problem, and then when they're in the Er they're admitted to the hospital. They're in the hospital for anywhere from a day to a few days. And you want to find when are those going to happen in the future?

LEONARD D'AVOLIO: Yeah.

DAVID SONTAG: What type of data is useful for that?

LEONARD D'AVOLIO: Sure. You don't have to just get through the ED, though. That's the most common, any unplanned acute admission.

DAVID SONTAG: Got it. So what kind of data is most useful for predicting that?

LEONARD D'AVOLIO: Yeah. So I think a philosophy that you all should take is whatever data you have, it should be your competitive advantage in solving the problem. And that's different in the way this has been done where folks have made an algorithm somewhere else, and then they're coming and telling you, hey, as long as you have claims data, then plug in my variables and I can help you.

Our approach-- and this is sort of derived from my interest from the start in solving the problem and try to make the tools work faster-- is whatever data you have, we will bring it in and consider it. What ultimately then wins is dependent on the problem. But you would not be surprised to learn that there is some value in claims data. You put labs up there. There's a lot of value in labs.

When it comes to behavioral health, and this is where you really have to understand health care, it's incredibly under diagnosed. There is a stigma attached to carrying diagnosis codes that would describe you as having mental health challenges. And so claims alone is not sufficient for that reason. We find a lot of lift from care management. So when you have a care manager, that care manager is assessing you and you are filling out forms and serving you and giving you different types of sort of functional assessments or activities of daily living assessments. That data turns out to be very powerful.

And then, a dark horse that most people aren't used to using, we get a lot of lift out of the clinicians whether it's the psychiatrist or care manager's notes. So there is value in the written descriptions of a nurse's or a care manager's impressions of what's wrong, what has been done, what hasn't been done, and so on.

DAVID SONTAG: So tell me a bit about the development process. So you figure out what you want to predict. You at least have that in words. You have your data in one place. Then what?

LEONARD D'AVOLIO: Yeah. Well, you wouldn't be surprised. The very first thing we do is just try to throw a logistic regression at it. We want the story to make sense to begin with, and we're always looking for the simplest solution to the problem. Then the team sort of iterates back and forth through based on how this data looks and the characteristics of it-- the density, the sparsity-- based on what we understand about this data, these guys are in and out of the plan. So we may have issues with data not existing in the time windows that you had described. Then they're working their way through algorithms and feature selection approaches that seem to fit for the data that we have.

DAVID SONTAG: But what error metrics do you optimize for?

LEONARD D'AVOLIO: You're going to have to ask them. It's been too long.

DAVID SONTAG: OK.

[LAUGHTER]

LEONARD D'AVOLIO: I'm 10 years out of being allowed to write code. But yeah, then it's an iterative process where we have to be-- this is a big deal. We have to be able to translate. We do positive predictive value, obviously. And I like the way you describe that, because a lot of folks that have been trained in statistics for medicine, whether it's epidemiology or the like, are always looking for an r squared or an area under ROC. And we have to help them understand that you can only care for so many people. So you don't really care what the area under ROC is for a population of, for this client, 300,000 in the one plan that we were serving. You really care about for the top 100 or 200, and really that number should be derived based on your capacity.

DAVID SONTAG: Yeah.

LEONARD So if I can give you 7 out of 10 for 100, you might go knock on their door. But for, let's say,

D'AVOLIO: between 1,000 and 2,000 that number goes down to 4 out of 10. Maybe you should go with a less expensive intervention. Huge education component, helping people understand what they're seeing and how to interpret it, and helping them connect it back to what they're going to do with it. And then I think probably, in courses to follow, you'll go into all of the challenges with interpretability and the like. But they all exist.

DAVID SONTAG: So tell me a bit about how it's deployed. So once you build a model, how do you get your client to start using it?

LEONARD D'AVOLIO: Yeah. So you don't start getting them ready when the model's ready. I've learned the hard way that's far too late to involve them in the process. And in fact, the one bullet you had up here that I didn't completely agree with was this idea that these approaches are easier to plug into a workflow. Putting a number into an electronic health record may be easier.

But when I think workflow, it's not just that the number appears at the right time. It's the culture of getting-- put it this way. These care managers have spent the last 20, 30 years learning who needs their help, and everything about their training and their experience is to care for the people that are most acute. All of the red flags are going off. And here comes a bunch of nerds and computer science people that are suggesting that no, rather than your intuition and experience of 30 years you should trust what a computer says to do.

DAVID SONTAG: So there are two parts I want to understand better.

LEONARD D'AVOLIO: Sure.

DAVID SONTAG: First, how you deal with that problem, and second, I actually am curious about the technical details. Do you give them predictions on a piece of paper? Do you use APIs?

LEONARD D'AVOLIO: Yeah. Well, let me answer the technical one first because it's a faster answer. You remember at the beginning of this, I said health care is pretty immature from a technical standpoint? So it's never a piece of paper, but it can be an Excel spreadsheet delivered via secure FTP once a month, because that's all they're able to take right now based on their state of affairs. It can be a real time call to an API.

What we learn to do informing a company serving health care is do not create a new interface. Do not create a new log in. Accommodate whatever workflow and systems they already have in place. So build for flexibility as opposed to giving them something else to log into. You have

very little time. And the other thing is clinicians hate their information technology. They love their phones, but they hate what their organization forces them to use. Now that may be a gross generalization, but I don't think it's too far off. Data is sort of a four letter word.

DAVID SONTAG: So over the last week, the students have been learning about things like FHIR and so on. Are these any of the APIs that you use?

LEONARD D'AVOLIO: No. So those are technologies with enormous potential. You put up a paper that described a risk stratification algorithm from 1984. That paper, I'm sure, was supported with evidence that it could make a big difference. I'm getting awfully close to standing on a soapbox again, but you have to understand that health care is paid for based on delivering care. And the more complex the care is, the more you get paid. And I'm not telling you this, I'm kind of sharing with them. You know that.

So the idea that a technology like FHIR would open up EHRs to allow people to just kind of drop things in or out, thereby taking away the monopoly that the electronic health records have-- these are tough investments for the electronic health record vendor to make. They're being forced by the federal government. And they saw the writing on the wall, so they're moving ahead. And there's great examples coming out of Children's, Ken Mandl and the like, where some progress has been made.

But I live in right now, I have to get this done inside of the health care of today. And very few of the organizations that we not just work with but would even talk to are in a position, like FHIR ready. In 5 years, I think I'll be telling you--

DAVID SONTAG: Hopefully something different, yeah. All right, so can you briefly answer that first question about what do you have to give around a prediction in order for it to be acted upon effectively?

LEONARD D'AVOLIO: Yes. So the very first thing you have to do is-- so we invite the clinical team to be part of the project from the very beginning. It's just really important. If you show up with a prediction, you've lost. They're part of the team. Remember, I say we're triangulating what they can and can't do, and what might matter what might not. They are literally part of the team. And as we're moving through, how would one evaluate whether or not this works? We show them, these are some of the people we found. Oh yeah, that makes sense. I know Mr. Smith. And so it's a real show and tell process from the start.

DAVID SONTAG: So once you get closer to that, after development phase has been done, then what?

LEONARD D'AVOLIO: After the development phase, if you've done a great job you get away from the show me what variable mattered on a per patient basis. So you can show folks the odds ratios on a model is easy enough to produce. You can show people these are the features that matter at the model level. Where this gets tougher is all of health care is used to Apgar scores which are based on 5 things. We all know what they are. And the machine learning results, the models that we have been talking about in behavioral health-- I think the model that we're using now is over 3,700 variables with at least a little bit of a contribution.

So how do you square up the culture of 5 to 7 variables? And in fact, I gave you the variables and you ran the hypothesis testing algorithm versus more of an inductive approach, where thousands of variables are actually contributing incrementally. And it's a double edged sword, because you could never show somebody 3,700 variables. But if you show them 3 or 4, then the answer is, well that's obvious. I knew that.

DAVID SONTAG: Right, like the impaired fasting glucose one.

LEONARD D'AVOLIO: Yes, exactly. So really, I just paid you to tell me that somebody who has been admitted is likely to readmit. You know, that's the challenge. So striking that balance between-- really, it's education more than anything, because I don't think that an algorithm created that uses 3,700 variables can then be turned into decision support where it can present you 2 or 3 that you could rely upon and then make informed decisions. And part of the education process is we also say forget about the number. If I were to give you this person, what would you do next? And the answer is always, well I would look at their chart.

The analogy we use that we find is helpful is this is GPS, right? GPS isn't going to give you like a magic, underground highway that we didn't know about. It's going to suggest the roads that you're familiar with. The advantage it has is that unlike you in the car as you're driving, it's just aware of more than you are and it can do the math a little bit faster than you can. And so it's going to give you a suggestion, and it's going to tell you more often than not, in your situation, I'm going to save you a few minutes.

DAVID SONTAG: Yeah.

LEONARD D'AVOLIO: Now you're still the driver. You could still decide to take 93 South and so be it. It could be that the GPS is not aware of the fact that you really like the view on Memorial Drive versus Storrow, and so you're going to do that. And so we try to help people understand that it just has access

to a little bit more than you do, and it's going to get you there a little bit faster.

DAVID SONTAG: All right, I'm going to stop you here because I want to leave some time for some questions from the audience. So I'll make the following request. Try to keep it to quick responses so we can get to as many questions as we can.

AUDIENCE: How much is there a worry that certain demographic groups are under diagnosed and have less access to care? And then, would have a lower risk edification, and then potentially be de-prioritized? How do you think about adjusting that?

LEONARD D'AVOLIO: Yeah, so that was a great question. I'll try to answer it very fast.

DAVID SONTAG: And could you repeat the question as quickly as possible as well?

[LAUGHTER]

LEONARD D'AVOLIO: Yeah. I mean, models can be biased by experience. And do you worry about smaller size populations being overlooked? Safe to say, is that fair?

DAVID SONTAG: And the question was also about the training data that you used.

LEONARD D'AVOLIO: Well, that's what I implied.

DAVID SONTAG: Yeah, OK.

LEONARD D'AVOLIO: OK. So all right, this work we're doing in behavioral health-- and we've done this in a few other environments-- if there is a different demographic for which you would do something different and they may be lost in the shuffle, we do bring that to their attention.

DAVID SONTAG: Next question! Is there someone in the back there?

LEONARD D'AVOLIO: You went too fast.

DAVID SONTAG: OK, over here.

AUDIENCE: How do you evaluate [INAUDIBLE]? Would you be willing to sacrifice the data of [INAUDIBLE] to re-approve the [INAUDIBLE]?

DAVID SONTAG: I'm going to repeat the question. You talked about how it's like reading tea leaves to just show a couple of the top features anyway from a linear model. So why not just get rid of all that interpretability altogether? Does that open the door to that possibility for you?

LEONARD D'AVOLIO: You're saying get rid of all the interpretability. I think the question was are you willing to trade performance for interpretability.

DAVID SONTAG: Yes.

LEONARD D'AVOLIO: And that could be an answer to it. Just throw it out. So if I can get our partners to the point where they truly understand what we're doing here and they have been part of evaluating the model, success is when they don't need to-- on a per patient, who needs my help basis-- see the 3,000 variables. But that does mean that as you're building the model, you will show them the patients. You will show them the variables. So that's what I try to walk them to.

DAVID SONTAG: So it's about building up trust as you go.

LEONARD D'AVOLIO: Absolutely. That being said in some situations, depending on whether it's clinically appropriate-- I mean, if I'm in the hundredth percentile here, but interpretability can get me pretty far, I'm willing to make that trade. And that's the difference. Don't fall in love with the hammer, right? Fall in love with building the home, and then you're easy enough to just swap it out.

DAVID SONTAG: Next question! Over there.

AUDIENCE: Yeah, how much time do you spend engaging with [INAUDIBLE] and physicians before staring to sort of build your model.

LEONARD D'AVOLIO: So actually, first we spend time with the CEO and the CFO and the CMO-- chief medical, chief executive, chief financial. Because if there isn't at least a 5 to 1 financial return for solving this problem, you will never make it all the way down the chain to doing something that matters. And so what I have learned is the math is fantastic. We can model all sorts of fun things. But if I can't figure out how it makes them or saves them-- we have like a $5 million mark, right? For the size of our company, if I can't help you make 5 million, I know you won't pay me.

So we start there. As soon as we have figured out that there is money to be made or saved in getting these folks the right care at the right time, then yes the clinicians are on the team. We have what's called a working group-- project manager, clinical lead, someone who's liaison to the data. We have a team and a communication structure that embeds the clinician. And we

have clinicians on the team.

DAVID SONTAG: I think you'll find in many different settings that's what it really takes to get machine learning implemented. You have to have working groups of administration, clinicians, users, and engineers, and others. Over here there's a question.

AUDIENCE: Actually, it's a question for both of you, so about the data connection. So I know as people, we try to connect all kinds of data to train the machine learning model. But when you have some preliminary model, can you have some insights to guide you to target certain data, so that you can know that this new information can be very informative for prediction tasks or even design data experiments?

DAVID SONTAG: So I'll repeat the question. Sometimes we don't already have the data we want. Could we use data driven approaches to find what data we should get?

LEONARD D'AVOLIO: So we're doing this right now. There's a popular thing in the medical industry. Everyone's really fired up about social determinants of health, and so that has been branded and marketed and sold. And so now customers are saying to us, well hey, do you have social determinants of health data? And that's interesting to me, because they've never looked at anything but claims. And now they're suggesting go buy a third party data set which may not add more value than simply having the zip code.

And we say of course, we can bring in new data. We bring in weather pattern. We bring in all kinds of funny data when the problem calls for it. That's the easy part. The real challenge is will it add value? Should we invest our time and energy in doing this? So if you've got all kinds of fantastic data, run with it and then see where you fall short. The data just doesn't tell you, now go out and get a different type of data. If the performance is low clinically and based on intuition, it makes sense that another data source may boost. Then we'll try it. If it's free, we'll try it quicker. If it costs money, we'll talk to the client about it.

DAVID SONTAG: For both of those, I'll give you my answer to that question. If you have a high dimensional enough starting place, often that can give you a hint of where to go next. So in the example I showed you there, even though obesity is very seldom coded in claims data, we saw that it still showed up as a useful feature, right? So that then hints to us, well maybe if we got higher quality obesity data it would be an even better model. And so sometimes you can use that type of trick. There is a question over here.

AUDIENCE: We use codes to [INAUDIBLE] by calculating how much the hospital will gain by limiting [INAUDIBLE]?

DAVID SONTAG: OK, so this is going to be the last question that we're going to end on. And it really has to do with one of evaluation and thinking about the impact of an intervention based on their predictions. How much does that causal effect show up in both the way that you formalize problems, then evaluate the effect of your predictions?

LEONARD D'AVOLIO: Yeah. So the most important thing to know is no customer will ever pay you for a positive predictive value. They don't care, right? They care about will you help them save or make money solving a problem. So cost effectiveness starts at the beginning. But the nice thing about a positive predictive value approach-- and there's so much literature that can tell you what does the average cost of certain things having happened.

So the very first part of any engagement for us is well, you guys are here. This is the cost of being there. If you improved by 10%, if we can get approval to that, then we start to model. And we say well look, of the top 100 people 70 of them are the right people. Multiply that by the potential cost. If you think you can prevent 10 of those terrible things from occurring, that's worth this much. So cost effectiveness data is at the start. It's in the modeling stage. And then at the end, we never show them how good we did at predicting. We show them the baseline. We say baseline activities outcomes-- where were you, what are you doing, and then did it make a difference. And the last part is always in dollars and cents, too.

DAVID SONTAG: Although Len didn't mention it here, he also does quite some work when trying to think through this causal effect. And we talked about how you use propensity matching, for example, in your work. We won't be able to get into that in today's discussion, but we'll come back to those questions when we talk about causal inference in a few weeks. That's all for today, thanks.

[APPLAUSE]