PROFESSOR: OK. So last time, we spent the hour and a half talking about classification methods and their use of genomic technologies in direct and close to direct, which we called indirect, applications of clinical medicine.

And we talked about three different types of class methods again, class exploration, class prediction, and class discovery. And I think we finished with a discussion about class prediction, which is a basic approach that's used to-- that starts with microarray data from a certain number of patients, builds a certain model, and then tries to predict something. Everybody who was here kind of remember that, vaguely at least?

And the basic steps of class prediction were, first, to choose a gene set, so choose a number of genes that's representative of the data and that divides the data into classes for which you're targeting your predictions to construct a function, a prediction function, that will be a mathematical function that you'll take a new sample with its expression data, and plug into the function, and get out an answer, and then determine a rule to try and then use the rule to classify one way or the other, and then, lastly, validation.

So I thought what we would do today is go through an example. So we went through all this. So I'm not sure we need this anymore. I thought what we would do today is go through an example. And the example is related to breast cancer and the paper from Nature 2002 that we talked about. So the issue of *New England Journal of Medicine* that was published today had a number of articles related to use of these classification methods in AML, in acute myelogenous leukemia.

AUDIENCE: [INAUDIBLE]

PROFESSOR: Well, there's two different papers. There's an editorial and there's a perspective written. Maybe we'll look at the perspective for a second. And the perspective is called "Microarrays in clinical investigations." And so the perspective is pretty bold in its views here.

And essentially, the last paragraph, the take home message-- so the take home message is that our usual thinking about biomarker discovery in clinical trials is about to change dramatically. In the future, clinical investigations will consist of small trials with a high density of data, precise patient stratification according to the expression profile, and highly tailored analysis of microarray data.

So they are looking at the articles in this issue of gene expression and AML. And they focus on identifying the relevant-- so looking through 7,000 genes or I think one study used 133 arrays with 13,000, another with 25,000-- looking at large number of genes and their expression in patients and identifying smaller numbers that are very relevant to disease. And that's what's meant by the term biomarkers here. And they're making a prediction that this is going to change the way clinical trials are being done dramatically. This seems very unlikely to me that this will--

AUDIENCE: I was really curious because when you reading the articles, at least glancingly, they seem very bread and butter, just classification and some prediction. And I was just curious how they [INAUDIBLE] the *New England Journal of Medicine* is very reticent to take anything with [INAUDIBLE]. And so that sort of struck me really.

PROFESSOR: Well, it's a learning process for everybody. There are things to learn from the breast cancer work that was done. And we'll go over the *Nature* paper. And that also was published in the *New England Journal of Medicine* as well. Let me give you-- let's see. I had this one for you. And this one is important because when you have [INAUDIBLE].

So let's-- I didn't have time to review the AML stuff that's accompanying the editorials and perspective in today's issue. I think it would be nice to go through that and look at the data directly instead of the claims in the editorial. But let's do that for the breast cancer work that was done.

So this is the paper. You might want to have it just so you have it here, because we're going to-- I thought rather than talking in the abstract as we did last time, to really go through concretely the use of the prediction method here, which is used quite similarly in one of these papers here.

So let's go through this *Nature* paper on gene expression profiling predicting clinical outcome of breast cancer. So in this paper, essentially what these authors did in the abstract, you can see that they used microarray analysis on primary breast tumors of 117 young patients and applied supervised classification to identify a gene expression signature strongly predictive of a short interval to distant metastasis.

So again this is-- again, I often get tripped up by viewing things in terms of the specific tool that was used. And really what they're doing is doing prediction here. Some of our different pigeonholes that I try to put things in, class, exploration, prediction, and discovery. This is prediction. Supervised classification is one aspect of it. But there are other parts of this to doing prediction.

But what they were interested in, essentially, is taking patients and making good or bad predictions of prognosis of outcome based on their gene expression. So they did a variety of things. But most of this paper is focused-- we look at, say, the second paragraph on the first page after the abstract.

An unsupervised hierarchical clustering algorithm allowed us to cluster the 98 tumors-- now we're down to 98 from 117, I'll come back to that-- on the basis of their similarities measured over approximately 5,000 significant genes. And that's figure 1A.

So here they have things sideways. They have the specimens as the rows and the genes as the columns in that top diagram. And so this is hierarchical clustering of this data set that consists of a big table. And the table again has 98 patient samples and 5,000 genes and expression levels that-- I might be short on the paper.

Why don't you guys look on that for now, yeah. Yeah, this paper is up on the site. OK, so this is the data set that they constructed.

AUDIENCE: I just have a question about figure 1A in particular. What's the utility of something like this in the paper? I mean they're obviously they're not looking at this-- like if I were to-- when they get their data and they see something like this, they feed into a computer that does all this analysis. So what's the purpose of showing us this rather than just kind of cutting to the chase of what their actual results are.

Should this be standard for when you're doing these types of studies or whatever?

PROFESSOR: I think they're pretty valuable diagrams because they do allow you-- this is the black and white, so there's the color version-- they do allow you to see that there are different patterns of expressions of genes and different subgroups.

AUDIENCE: I mean, I guess maybe these lower ones are blown up, so where [INAUDIBLE] it may become more obvious. But I'm thinking just in terms of if I saw an SDS-PAGE gel in the paper, it's just like, right there you can see the results. If I look at something like this, it just looks like a bunch of pixels [INAUDIBLE]. I just I guess.

PROFESSOR: So you're asking about the motivation?

AUDIENCE: Well, yeah, I mean so--

[INTERPOSING VOICES]

AUDIENCE: Yeah, so like the sub figures can be useful. I guess and maybe I can see how you'd say, OK, this is the reason we expanded on it.

PROFESSOR: But even in the big figures, you can often see, and maybe I have a picture of this on color that I can put up, I'm not sure if I do. But I should. But you can see that there are different groups, even within color. [BEEPING] Sorry, I have some patients.

AUDIENCE: I guess sort of as a newbie here, looking at my [INAUDIBLE] traditionally, you see something like that-- when I'm seeing something like this, I just go, well, gee, that looks pretty. And I have no idea what it means or whatever. And I think that would be kind of intimidating for people to want to pick up a paper like that. And they don't quite get the point of seeing something like.

- Well, there's two issues. That's an independent issue. I think there's an issue that I find talking about these things is, the same thing as looking at a scattergram, why should you involve a-- look at a scattergram rather than just the correlation coefficient described in here. Just give me a look right path to see, to actually see, by visual inspection that it looks like this class is there.

And so when I look at this thing, I can actually convince myself but there's a lot of crap. But there is also some big chunks of red.

AUDIENCE: So definitely the color picture helps a lot.

PROFESSOR: The other part of these pictures, which is helpful, are the dendrograms. OK, so again we talked about the use of hierarchical clustering as a classification method and that it does not really produce classes. OK, again, everything is related to everything else. You don't have disjoint classes. It's just that the degree of relationship is represented by the distance, literally the length of this line, right here in the diagram, represents the relationship between these two over here and say these two over here.

And so again everything every cluster is split. This is binary. So everything is split into two. And so the adjacency is actually another issue too, about these, is that you can spin any one of these clusters at any node without changing the organization here. So sometimes you will see diagrams-- why don't I show you one since we're getting to nitty gritty here.

AUDIENCE: When you have a paper and you see that figure, just apart from of the obvious, like here's a really big red patch, [INAUDIBLE] that's-- you're also, if someone is familiar with this, you're also gathering other information. By not having data in hand, you can see, you can look at that figure and say, ah, pattern, but here's another part of the figure that also gave me some more information.

And me as the newbie, I'm just not seeing it because I'm not as familiar with that.

PROFESSOR: And I don't have a password here. I can't do this.

AUDIENCE: Is that a fair assessment?

PROFESSOR: That's a fair assessment, OK. What I was going to show you is that there's been, in my area, last week there was a publication-- it's the second one this group has done-- where they put one of these diagrams up here. And they're showing that three patients with one condition, they all cluster together and separate from another seven.

But they don't put the dendrogram on that. All they show is adjacency of rows. And they don't show exactly how they're organized or connected. So you see three that are here and then another seven that are here. But it's possible that this one over here is actually part of a group of four over here.

And so that's another value to these diagrams, is when people claim classes to look and see. When you look at this one, it's not overwhelming, actually. And this is where they drew the line. And they took this as one class. But this is really-- I mean every one of these is just subdivided here, instead of one fairly large class. So there is helpful information for seeing those diagrams.

AUDIENCE: So the leftmost is the two classes in [INAUDIBLE] 117--

PROFESSOR: That's correct. They said they are two classes. So this is one branch here. And this is the other branch here. OK. OK, yeah, well, let's get into the nitty gritty, at least so we have one example that we've understood fairly well what's happening.

OK, so this is an example. So this is just hierarchical clustering. What would you call this if I made you pigeonhole it into class exploration, class prediction, or class discovery? These are 98 patients with breast cancer. These are both lymph node positive and lymph node negative. They're pretty much all of their samples except for a group that they've left off for other purposes.

This would be an example of class exploration. So they're starting by looking at their data. And they're just saying, well, does it look like the expression profiles differ among patients and allow us to find different subgroups, different profiles?

And we talked about one step that was important to do before this. And that was the stuff about signatures and their reproducibility and their distinctness. They don't provide any data here in that regard. They say that they did every tumor sample twice, that they did two independent array profiles on each one. But they don't say anything about what they did with that data. They don't provide any correlation coefficients, for example, to say that when you do the same experiment twice on a given tissue, you get out the same pattern of genes.

But accepting that, this first step is exploration. And here there is an attempt to correlate these groups with different clinical features. So first, they do hierarchical clustering of the entire data set. And they look and see broad patterns. Maybe there is a group over here that's more red for these genes and greens for these. And so maybe there is some structure here.

So the next step they do is to, for each one of these patients, they look at binary outcomes, and say if this patient here had a mutation in the BRCA1 gene, they'll color it in black. And this patient here did not have a mutation.

So for each one of these 98 patients, they're now looking at one, two, three, four, five, six different phenotypic markers and coloring it in this way so that we can then look and say, is there any relationship between these phenotypic variables and these two classes which were defined by the dendrogram?

Again to pigeonhole it into one of the three class methods, what are they doing here? What's that? Class exploration, class discovery-- they're doing class discovery here. So they looked at their data. They explored it. They wondered that there might be classes. And they're now asking is there something significant phenotypically, something meaningful clinically, that relates to these two classes? Or is it just an artifact of doing clustering, which will always reduce things into classes?

Depending on where you cut this tree, you can have two classes here, or this one here that looks like there's a large class, which goes from here down to here. You can make up classes at any different level here, because fundamentally there really are no classes with a dendrogram.

But they're now saying, well, what phenotypic variables might correlate with this split right here? This is still discovery. Yeah, I would call this class discovery, that they found two classes. And now that to establish them as truly classes that have some meaning, some clinical meaning, they want to say what clinical variables might this relate to here?

And so looking at the things here, there are few things which are fairly striking, I guess. The presence of estrogen receptors on the tumor cells here is pretty good. Most of these patients are ER positive. Most of these patients here are-- actually, I can't remember which is black and which is white, which is positive and which is negative for their classes here.

Metastases-- so the presence of-- and I think this is distance metastases at five years-- not really too much difference. Maybe I think white is present. So maybe more of these. What's that? White is positive. So there may be more consistently, certainly a larger percentage of this group, had distant metastases at five years.

So once they do this and believe that they do have classes here, they then move to the next step and say we're going to build a predictive model. And this is the one that they focus on, is whether there's metastases or not. So they're saying in this data, there seems to be some way to classify the data that has some distinctions in terms of predicting outcome. And we're going to take it a step further here and actually go through the trouble of building a model to predict this.

AUDIENCE: So they want to use the metastases as the principal predictor for or [INAUDIBLE]. Why, maybe this is just another ignorant question.

PROFESSOR: No, there's no ignorant questions here.

AUDIENCE: Why did they choose that as like estrogen receptor?

PROFESSOR: Well, yeah, because people already know about the estrogen receptor and BRCA1 mutations. OK. So and metastasis is the-- it's fundamentally the important thing. It's the important clinical endpoint. If we can make better predictions based on gene expression profiles of tumor when we do an initial biopsy on the breast cancer, can we use that to then decide how to treat people?

People have a good prognosis don't need the same treatment as people with a bad prognosis, maybe. It doesn't actually-- there's no treatment intervention-- not a treatment study here. It's not saying that we can intervene and alter prognosis through treatment. It's strictly predictive of prognosis at this point. OK?

OK, so let's go back to the paper. And so the next thing they do after this class exploration, this sort of partial class discovery, I mean they don't carry it further and say that we really have discovered a new class here, because they really haven't. I mean what the prognosis is really correlating with is the estrogen receptor status here, which is a well-known fact.

So patients who are estrogen receptor positive have a better outcome, I think, in breast cancer. So I'm not sure about the black being positive or negative. The black is negative, OK. Yeah, that's right. So patients who are estrogen receptor positive have a better outcome. The estrogen receptor negative patients, by and large, have a bad outcome.

So there's actually nothing new. There's no important class discovery that took place in this paper. So what they next do is change the data set. Let's shut this off. It's not shut down. So the next step that they do in this paper, and that's on this page 532. Bottom of the first page, sorry, bottom of page 530-- is they focus now on a subgroup of these samples on the 78 patients with sporadic lymph node negative disease.

So these are all patients who, at the time of the diagnosis of their breast cancer, had lymph node biopsies which were negative, and did not show any spread to the lymph nodes. And so they then focused the rest of the paper on this select group of patients, lymph node negative, and also sporadic, which means none of these had mutations, the BRCA1 or even BRCA2 mutations. They were not genetic breast cancers, patients who are genetic.

And so in that 78 group, 44 patients were free of disease after five years, and 34 patients developed metastasis. So we're down to 78 patients. And 44 had good prognosis, or turned out to have good outcome. And 34 had poor outcomes.

So what they then say on the next page, 532, is to identify reliably good and poor prognostic tumors, we used a powerful three-step supervised classification method similar to those used previously. In brief, approximately 5,000 genes were selected from the 25,000 genes on the microarray.

OK, so let's get into how they build the model here. So they're going to use this data. So they're going to use 78 instead of 98. And they do have 5,000 genes. So the microarray has actually measured 25,000 genes. So they had this data set to start with.

And the first thing they did is use basic, but still fairly ad hoc, procedures to get rid of genes that had very little meaning, so genes that just didn't vary, a gene that had a low expression level for every single one of these. And they removed such genes and ended up with 5,000 after that point.

AUDIENCE: So was 5,000 predetermined, or was that--

PROFESSOR: They-- let's see how they ended up with five. They said significantly regulated and more than three tumors out of 78. So this is an area which remains completely ad hoc and different in virtually every paper. We call it the initial filtering of meaningless genes. And they decided that-- I believe that they did T-tests for significance and required there was some significance in at least three out of 78 at some particular significant level, and ended up with 5,000 genes.

There's many ways people do this. Some people will say I'll look at the standard deviation of these gene vectors. And if the standard deviation is less than some cutoff, I'll assume those genes are not varying enough in this data set and get rid of those.

So now they want to build their predictor. And so we talked about those three steps. And the first step is choosing a discriminative gene set. So they're going to whittle the 5,000 genes down into small numbers, which they're going to make put in their black box to make their model. And they do this by correlation with ideal outcome. I think we talked about that method before.

So first they'll arrange the data set so that all of the good prognosis are up here. OK, and then T+ 45 through 78 are the poor prognosis. And then they'll make up an ideal vector, i, and put let's say 1's here and then 0's here.

And then for each of these 5,000 genes, they'll calculate a correlation coefficient to this ideal vector. So r for gene one correlated with ideal vector i was some number, 0.6. And r for gene two correlated with vector i is some other number, negative 0.45, let's say. And do this for the 5,000 genes. Everyone follow?

So the first thing they do is they get this down to 231 genes by using a cutoff of 0.3 negative or positive. So anything that was between an r between 0.3 and negative 0.3 is gotten rid of here. So they get rid of this one here. This they would keep. This one they would keep, and so forth.

And so that left them with 231 genes, which we'll renumber and call these genes now. Yes--

AUDIENCE: Question related to the coefficient for [INAUDIBLE] normalizing their variables in [INAUDIBLE] gene expression so as to be in the same order of the outcome variables, presumably binary. What kind of transformation goes on?

PROFESSOR: Well, there's two issues here. One is normalizing the data set. So they used cDNA arrays. So their normalization was built in because they used a reference. So a cDNA array is an experiment you have two samples applied each time. And their reference consisted of a little bit of RNA from each one of, I think, the larger set. I don't if it's the 78 or the 98 or 117 samples. So that was the normalization.

To do this, they just organize it that way. Just move these up here and move these down here.

AUDIENCE: I guess my question was, I'm just trying to figure out the [INAUDIBLE] do you actually quantify for-- in the correlation coefficient, you're looking at direction. But does the amount of expression or difference in expression between positive and negative side matter In terms of your correlation?

PROFESSOR: No, that won't matter. Again, you could-- again, Pearson correlation coefficients, at least, are invariant to linear transformations. So you could take every piece here and multiply by 15 and add 7, and it won't change its correlation coefficient.

OK, so now they were left with 231 genes. And all of these genes have some threshold of high correlation, either positive or negative to this ideal outcome here. So gene one had a certain set of expression values in this group. And it was different than in this group, because it correlates, has a high correlation coefficient to that factor. Everyone follow? OK.

OK, so now they do-- some people would stop here and say this is my discriminative gene set. And the next thing I'm going to do is build my prediction function. And then I'm going to make my rule. And then I'm going to validate.

But they decide to further optimize this discriminative gene set. So this is a little complex. But what they do-- not that complex. So they take the 231 genes that they have here. And they rank order them by largest magnitude of their correlation coefficients.

So gene 16 had a correlation coefficient of 0.9. And G 12 had a correlation coefficient of negative 0.85. And they rank these going down to G 231, which just made it with a correlation coefficient of 0.31, into this group.

So now they're even looking deeper into this data structure. And they're looking for the very best genes that correlate with this outcome.

An example of such a gene might be gene 16, which had expression levels that were 1,000 for these samples, and was 5 for all of these samples. It was a perfect classifier, that one gene.

So now they're going to really build a discriminative gene set through the following procedure. They're going to take the top five from this list and call that our discriminative gene set consisting of only five genes. OK, then they're going to go through this, build a prediction function, make a rule, and validate using leave one out cross validation here.

Then they will add five more genes to this discriminative gene set. So they'll go down five more in this group and now have a discriminative gene set of 10 genes. They'll take that, build a prediction function, make a rule, and validate again. Now that they've done a second validation, they have a second accuracy number to look at.

And they'll say did my accuracy get better or not? So the first time they did this with the top five genes, they found an accuracy in terms of prediction of, I don't know what it was. But we'll make up a number. Let's say 60% predictions were correct after they built their model.

So then they did this with another of the larger gene set, including the next five best genes. And they found they had a prediction function of 70, that got them up to 70%. And they kept on going until their accuracy was the best it could possibly be and wasn't getting any better with the addition of more genes from the bottom of this list. There they stopped at 70 genes.

So they constructed their discriminative gene set right here as a set of 70 genes based on this optimization of trying first the top five, building the model, testing the model, and looking at its accuracy, and improving the accuracy until it was as best as it could be. Yes--

AUDIENCE: Why five and not one?

PROFESSOR: One would have taken them five times as long to get there.

AUDIENCE: That's true but [INAUDIBLE] one is more relevant.

PROFESSOR: Well, that's actually the point that people are trying to make in this field. That one is not relevant, in that the biomarkers that we're now looking for are groups of genes.

AUDIENCE: I mean in the sense of implementing, I don't mean--

PROFESSOR: This can be implemented clinically. Once you set up to make an inkjet-synthesized or even robotically spotted array, you can do it for a few dollars. You can spot arrays for a few dollars, basically. This is not actually expensive.

AUDIENCE: So basically they're going to find the minimal amount of these genes, and we give them the maximum amount of predictive value [INAUDIBLE].

PROFESSOR: Correct, a predictive accuracy in terms of going for this one prediction of good versus bad outcome.

AUDIENCE: Is there any value-- so at each step before they got to the extra revision that you mentioned, [INAUDIBLE] close where most people would stop. Is there any value in there going outside of their 231 back into the 5,000. And just randomly--

PROFESSOR: I don't know how far up they went. I don't if they went past 231. I doubt it, that they then looked back at other genes.

AUDIENCE: [INAUDIBLE] and they never [INAUDIBLE] at 231 genes, then they said OK, well, from there we've got certain candidates, which I think is quite fair. Once they've got their 70, is there any value in going back out to the larger data and seeing if we can call a few more things that maybe--

PROFESSOR: Maybe, but the main critical point of this approach is that this is overfitting, this is very, very serious overfitting of data. I don't think this is what you want to do. But people are still doing it. And I mean, it's fairly reasonable to make this gene set here and stop there.

But to do a procedure that you then optimize this set, you stop with the 231. To then take that and to do repetitive cross validation on the very data that you're using to build the model, and to do it to select the genes, even in that way, is very serious overfitting of a data set.

And I put up that fictitious graph yesterday about the points and overfitting it with the best possible curve. But it may not be the right one to make a prediction of a new point here because of how overfitting it is. So it works perfectly for the data set.

And I think I brought up some of the pitfalls in terms of validation that's been done with microarray data. I think I passed out a copy of that paper mostly.

AUDIENCE: Can't be overcome by cutting the data set in half and doing exactly what you just described, the method that you just described, and then taking that-- [INAUDIBLE] and then taking that other half a [INAUDIBLE] because you did describe--

PROFESSOR: You can build-- right, that's the right way to do it. So that's the training set and then the validation set. I mean even the real right way to do it is to make up any model you want and then prospectively test it on the next 100 patients.

AUDIENCE: [INAUDIBLE] is that typically, the real performance is particularly in 2004, you have so few patients that you loathe to get away from many of your test set, because you might not have enough signal in your training set that you could [INAUDIBLE] for people. Otherwise, [INAUDIBLE]. That's why it's so amazing that [INAUDIBLE] can [? write ?] a paper besides a third of patients [INAUDIBLE].

PROFESSOR: Yeah, when these chips first came out, they were $2,000 a piece. Yeah, they're very-- they're even more than that, but--

AUDIENCE: Even more [INAUDIBLE]

PROFESSOR: So these were very costly experiments to do back then. So this is their discriminative gene set. OK, now they have to choose a predictive function. So the predictive function, again, is if one was to give this model a new clinical specimen, how do we then make a prediction on the new clinical specimen?

And so they used a fairly simple predictive function here, which is correlation coefficient to this ideal outcome. So if I gave you a patient tissue sample. And you did one of these microarray experiments and measured 25,000 different genes on it, then either you would use that in this model by pulling out the numbers from the 70 genes that were relevant-- that were part of the discriminative gene set. And then taking that number of 70 genes and-- let's see, sorry, I'm goofing up here. So-- improved--

The classifier predicted correctly the outcome. I'm sorry, I thought I understood their predictive function in this paper. Let's take a peek at it. The predictive function is a threshold rule. And that's evident like on figure 2.

I believe they took an average number. So they took the average. So bad profile-- they took these profiles here and just average. So there's 34 in this bad group, bad prognosis group. And they took the 34 numbers here and averaged them together to get one number here. And then they did that for their other genes. So we're down to 70 genes right now.

And so they have this ideal bad profile. And then pretty sure this is what they did. And they did the same thing for the good here. And they have their good profile, which is just an arithmetic mean of the expression levels.

And now when the model is presented with a new row vector there of expression data for 70 genes for patients, and we'll call that new patient patient n, then they will calculate the correlation coefficient of patient n to maybe just the good vector. I guess you can't do it.

Well, so I'm sorry I'm forgetting the details here a little bit. And I'm having trouble finding them right off here. But I'm pretty sure this is what they did. But so then which one did they correlate it with? The good one? Uh-huh. OK, yeah, which brings up another point, why not do it to the bad profile and correlate to that as a means of your algorithm?

That might give you a very different answer. I'm not quite sure. But in any event-- so, OK, it seems like there is this choice here that they look at its correlation coefficient to the average good profile and got a number. So this will give you r, we'll give you a number, 0.4. And that's their prediction function.

And then the last step in this model building is to make up a rule and say if there's-- and they used a threshold rule and a non-ambiguous classifier. So if it's above the threshold, it's in one group. If it's below the threshold, it's in another group. And they actually explored a couple of different possibilities for the threshold.

And we talked about that Tuesday that this rule and varying the threshold is the classic trade off of sensitivity versus specificity here of a test. And they talk about that in this paper and show-- so on this diagram here, so how to divide the blacks and the whites to best advantage is the question here, the good versus the bad prognosis. And do we cut it here and get all of the whites with just a couple of blacks or do we cut a little bit more and get more whites? But we're trading off sensitivity and specificity here. And that's the and that's the rule that they use.

So that's the model that they use. And then they actually do a couple-- in this paper, they do a couple of things for validation of it, for this step right here. And they do pick out a new test, a new validation set of 19 samples that they did not use to build the model here. And it's small. They don't say exactly how they constructed that group of 19.

But they do use a separate group of 19 to then test this on. And when they do that-- so that's I guess page 534, the third paragraph-- to validate the prognosis classifier and additional independent set of primary tumors from 19 young lymph node negative patients were selected. There were seven patients who had the good outcome, 12 who had the bad. And it resulted in two out of 19 incorrect classifications.

I think both-- so that's how they got an accuracy number out of that, two of 19 as the accuracy. One of the important issues that you get into when you do something like that is the accuracy may not be the same for the good versus the bad patients.

So you might have a classifier that predicts all of the good outcomes correctly, but is terrible at predicting the bad ones. Half the bad ones it says are good outcomes. And that's not very helpful to you. And so I think the data is in here. And I think it's mentioned in the pitfall article that I handed out. That teases that out a little bit more as to the importance of saying whether the predictor is working equally well on the good and the bad cases that are presented to it.

So I think that's essentially this paper. There was a follow up for this, which was in the *New England Journal.* It's on the website, which is even more focused on prediction as survival, gene expression signature as a predictor of survival in breast cancer.

They used a larger data set here. But they use the same model here. They did validation on the data set. But they didn't do the full validation procedure. So we talked about that, how a lot of validation is done through leave one out. So leave one sample out, build the model, and then test that sample. And then do it for each of the samples 78 times or 200 times, whatever. And see how accurate your predictions are each time.

But it's important when you do that to go back and rebuild the choice of the discriminative gene set each time. And this actually is in the *New England Journal of Medicine.* They do actually report both. They're aware of the need to do this. And the numbers I think were 24% versus 41% or something, 27%.

So when they do the error-- when they do the validation based on not doing this step, here we're choosing the gene set, there was a 27% error rate there. But when they did it repeating this gene step, there was a 41% error rate, which is getting close to chance, flipping a coin.

But in any event, so really on the basis of this, the Netherlands Cancer Institute that did this said that they were going to start using this 70 gene set to make clinical decisions. They're going to take new patients with breast cancer. Do a microarray experiment. Measure the 70 genes. Use this model, and make a prediction about good or bad outcome, and tailor their treatment according to that. Yes--

AUDIENCE: I actually had a question related to this and more [INAUDIBLE]. So the lab puts out a study like this, says here's 70 genes that we think are good indicator of [INAUDIBLE] [? By the logic, ?] I want to validate these results, right? Now in gathering and collecting this data, like this ridiculous study, they're looking at patients over five years.

You do a microarray analysis on a patient and they pop up positive or whatever for this indicator, now you've got people making clinical decisions on this. And it seems like the validation method is an [INAUDIBLE] withheld data. They need to have a bad prognosis or a bad outcome in order to say that these candidates-- for the patient-- if these 70 genes actually do have a negative effect, or indicate [INAUDIBLE] that effect has to be there.

So how does that how does a clinician or a clinical researcher, who kind of sees patients and does this, how do they balance that? I mean what would your suggestion or advice be? Does that makes any sense at all?

PROFESSOR: Well you're talking about how do you do research on patients this area?

[INTERPOSING VOICES]

AUDIENCE: [INAUDIBLE] how you [INAUDIBLE] patients clinically?

AUDIENCE: Well, [INAUDIBLE] so we have this result, people need to validate it. In the process of validating, they're going to find people who have this particular pattern, right? And they need to follow those people over a certain time course. And in order to validate their own [INAUDIBLE] essentially, if none of the people who have these 70 genes, or expression profiles or whatever, develop breast cancer, then the model's kind of [INAUDIBLE] back to the drawing board.

So essentially, and I hate to say it this way--

PROFESSOR: No, say what you're thinking.

AUDIENCE: They need these people, in order to validate the model, they need them to progress right to disease state. And yet at the same time, they're, as clinicians, so putting aside the researcher part, as clinicians, they need to treat these patients. And if there's a strong indication from prior studies that the chances are you're going to get breast cancer and it's going to metastasize. And it's going to get really bad for you. They can't just not treat these people aggressively. Or that treatment is going to affect that outcome.

PROFESSOR: Well, people-- yeah, yeah, so I don't think anybody's doing--

AUDIENCE: You think about the Tuskegee-like experiments.

PROFESSOR: Yes, so nobody's denying treatment or withholding best known treatments to patients.

AUDIENCE: So how do you calculate that into the validation of the model? That if you get-- you can't-- because you can't say, well, they didn't have metastasis because of my treatment.

PROFESSOR: There are a few issues. And you actually-- so in this *New England Journal of Medicine* paper-- here, I did not pass that one out. It's on the website. What they did is they used this model in a larger group of patients and they then compared predictions of this model to best clinical predictors.

So for example, there's other predictive scales. And they used one called the St. Gallen scale, which predicts outcome in breast cancer. And they said how does our model compare to the St. Gallen scale. If we take all of these patients and we see what their St. Gallen number was and how they did. And we take our predictions and how they did, which is better?

And they concluded that their method was better than the St. Gallen method of making predictions. And that was really one of the principal-- the principal justification for saying we have a better method to make predictions about how patients are going to do. And we're going to start using this in the Netherlands at the Cancer Institute to make our decisions, rather than St. Gallen.

And what you point out is actually a big flaw in research that does this, because patients had their St. Gallen scale determined when they were diagnosed. And depending on what their clinicians thought, treated them accordingly, in order to achieve best outcome. So if they were in a poor outcome group to begin with, based on St. Gallen criteria, they got treated more aggressively to try and make their outcome better.

So actually, the goal of the clinician is for the St. Gallen scale not to be right or this group of patients. The interventions that took place on these patients, because it was not a perspective controlled study, the interventions that took place were specifically designed to mitigate the poor prognosis that the St. Gallen criteria indicated on these patients.

And so it may not be a surprise that the St. Gallen predictor was not all that accurate later on, because people were intervening actively with the results of that predictor to change outcome. And that wasn't happening with gene expression profiles. Nobody looked at the gene expression profile of these patients 10 years ago and then decided how they were going to treat them over the next five years to make their disease better.

AUDIENCE: So is it just that because for right now the turnaround on the analysis of the gene expression profiles is long enough that the St. Gallen approach of aggressive treatment or whatever, that turnaround is long enough that you'll have your answer and your validation before clinical decision start getting made, repeat experiments are done to cross validate--

PROFESSOR: They'll never be a-- I mean to get a new drug approved in this country for labeling requires that you take a group of patients with a disease, and you treat some of them with the drug and some of them with a placebo, and you see which works.

To ever really know whether prediction models are going to work accurately, one needs to do the same thing, to take patients prospectively to apply prediction models to them ahead of time, and to see what happens with them, without an intervention that's based on the results of the prediction.

That's why I find a paper like this perspective here kind of scary. The microarrays in clinical investigations from today's journal, where the authors envision that in the future clinical investigations will consist of small trials with a high density of data, precise patient stratification according to expression profile, and highly tailored analysis of microarray data, otherwise known as massive overfitting of tons of expression data in a small number of patients here.

And that's never going to tell you whether your predictions are correct. And if you're going to make decisions about treatment or predictions based on that, you won't be accurate. I don't believe that there'll be sufficient accuracy from small. I don't think that trials need to get smaller. And trials can be made smaller because we now have microarray data. In fact, trials need to be probably larger, because you have that much more undetermined data set.

AUDIENCE: If you're [INAUDIBLE] the reproducibility [INAUDIBLE]

PROFESSOR: Well, so the Netherlands Cancer Institute, they patented their set of genes as a predictor of breast cancer. So, hoping-- right, well, hoping that they'll have the patent. And if they show that their stuff works, then people will come to them and they'll start wholesaling their type of arrays and their approach.

Do you know that-- I haven't seen anything, Zack. I mean there was an article that they were going to start doing this in the summer of 2003. And it was in *Nature.* We're going to start doing this. And silence. Did they come to their senses?

AUDIENCE: That's what I'm thinking must have happened. I think that probably somebody who actually knew this [INAUDIBLE] say, do you want to screw our patients over by doing this? I don't know that for a fact.

PROFESSOR: Right.

AUDIENCE: Yeah, what's an article [INAUDIBLE]

AUDIENCE: Oh, yeah.

AUDIENCE: I can't remember how many marketing [INAUDIBLE]

[INTERPOSING VOICES]

AUDIENCE: There's two issues here. One is whether FDA approves these gene chips in general. We'll get back to that in a second. Narrowly, whether you use a [INAUDIBLE] formula to prescribe these kind of [INAUDIBLE] on any platform. And I think as a responsible clinician, most of us would not right now necessarily stratify our patients based on the prognosis alone. You agree with that?

PROFESSOR: Yes.

AUDIENCE: It's been done kind of for cancer. And some oncology trials, it's kind of scary to me that have been done. [INAUDIBLE] been conducted. The way cancer is done in this country is within impactive centers and cooperative oncology groups where large numbers of individuals are treated in protocols. And those are now being stratified by [INAUDIBLE] for some studies.

I can't say that I fully agree with that. [INAUDIBLE] actually using the measurement [? technology. ?] Roche Diagnostics made the headlines about two months ago when they tried to get a chip, not for expression, but for genotyping mutations in the P450 of the proteins that clear idea of toxic chemicals out of our body in the liver.

They tried [INAUDIBLE] it's very important for [? promo ?] genomics because it'll tell you how fast growth appeared. Roche tried to get it simply through a waiver that essentially said this is just a method of measurement. We don't really have to get specific FDA approval. The FDA said no, no, no, hold it. Let's go through the formal approval process.

So what Cecily is talking about is that it announced that another [INAUDIBLE] will see if they can repeat it the same way. And this is-- It's actually an act of valor. As some of you are involved [? in HSP ?] should know that HSP is actually very much involved in help the FDA figure out exactly how to look at this, because [INAUDIBLE] the FDA has only now developed the minimum standards of data submission from the pharmaceutical companies on the microarray data. They have no guidelines whatsoever on the standard identification patients.

So I think there's going to be several hops, skips, and jumps to get these kind of measurements translated from research descriptions to clinical measures.

It's still going to be used, by the way, to give you a-- it still can be used in a Medical Center without FDA approval as long as they're not sold. So for instance [INAUDIBLE] the genetics and genomics has contracted with Affymetrics for a sequencing chip for genes involved in hearing loss and cardiomyopathy. But [INAUDIBLE] is not selling this for general use. It's been used internally by its clinicians.

So it's a very tricky process that that's been where regulation-- it's genomics in fact [INAUDIBLE] will be giving a lecture about that [? later on. ?]

PROFESSOR: Is Margulies giving that lecture?

AUDIENCE: Yes.

PROFESSOR: OK, that's all for today.