ARMAN KILIC: All right, my name is Arman Kilic. I'm going to be giving a talk today on artificial intelligence and machine learning in cardiovascular health care. AI and machine learning is a hugely popular topic, and we hear about it in the media almost every day. Here are a couple magazine covers looking at the Economist, Science, Forbes, Time Magazine, all of which have had AI and machine learning as major features in their articles, and we're hearing about this more and more every day. In the health care market the artificial intelligence market is expected to grow exponentially. Currently it's estimated to be about $2.1 billion, and by 2025 it's estimated to grow to over $36 billion as a market in AI.

A couple definitions that I think are worth going over. There's a lot of misconceptions about what this means. A lot of people think these are machines that are going to be taking over the world. Artificial intelligence is really a broader term and it refers to technologies or systems being able to demonstrate human like intelligence. Machine learning is a subset of artificial intelligence, and it refers to the ability of machines to learn from data to improve at tasks with experience, and to make predictions.

There are various types of machine learning algorithms, and generally they're classified into two categories. So one is supervised learning and the other would be unsupervised learning. Supervised learning is where you have known inputs and known outputs, and the goal of that algorithm is to accurately map the inputs to the outputs. Unsupervised learning, there are no labeled outputs. There's only inputs. And essentially what you're trying to do is learn more about the inherent structure of the data that you're working with.

Here are some examples of machine learning algorithms, and we'll go into a few of these. Supervised learning, again, where you're trying to map inputs to outputs. These include things like decision tree analysis, neural networks, extreme gradient boosting, support vector machines or random forest. There are several examples of unsupervised learning algorithms as well. Again, with these you're just trying to learn about the inherent structure of the data set that you're working with. This includes k-means clustering, dimensionality reduction and hierarchical clustering.

So just to give a general overview of some of these algorithms. This is an example of an unsupervised algorithm, which is called k-means clustering. So essentially the user that's working with the data will define k, which refers to the number of centroids in the data set. The centroid is essentially a center of a cluster of data points within the database. And I'll show a visual depiction of this in a few minutes. And then each data point in the data set is then assigned to one of these clusters to basically minimize the amount of distance between that data point and the defined centroid, or the defined center of the cluster.

And the centroid is essentially the way the algorithm works, the centroid are initially randomly selected. It's an iterative process where they're mapped to different locations and iterative calculations are performed, and eventually we figure out where the optimal place of these centroid or these clusters of data are. I think this helps explain that a little bit better. So each of these clusters are depicted by a different color in this diagram. So we have a y plane, we have an x plane. You have a cluster of data that are defined as green that seem to be similar to each other. You have a cluster of data that are blue and a cluster that are red that are very close in proximity to each other, so these are the clusters of data that the algorithm has defined.

And each of the asterisks that we see and these represent a centroid, or represent the center of each cluster. So this gives us some insight into the inherent structure of this data set. It's able-- it's allowing us to group various data points according to similar characteristics.

Another example of an unsupervised algorithm is called dimensionality reduction. So what this tries to do is really simplify the data set that you're working with. So if you're working with a data set that has 100 variables in it, and you want to simplify it, dimensionality reduction allows you to really minimize that and reduce it to just a few features or a few dimensions.

The real benefit of dimensionality reduction is it helps address the problem of over fitting. So over fitting is a term that's used very frequently in statistical analysis and in particular of machine learning, and it refers to when models have really-- they're too complex. They have too many features, and it makes them increasingly dependent on the data on which they're formed. And what ends up happening in the case of over fitting is that these data, or these models tend to perform very poorly when you use external data to validate the models. The real benefit, again, of dimensionality reduction, you really shrink the size of the data set you're working with. That reduces computing time, reduces computing power and helps with storage space and things like that.

This is an example of a supervised algorithm. So in these cases, what we're trying to do is we have inputs, we have outputs, we're basically trying to create models that will determine how well we can predict that output. So decision tree analysis is basically, as the name suggests, a tree. It's a decision tree. And there are conditions or there are internal nodes that are used that then split into branches of the tree. And when you have a branch of the tree that no longer splits, that's technically the leaf for the decision of that algorithm.

There are important aspects to know when you're using this algorithm. One is you need to know the conditions upon which you're going to split, right? So this could be something like age. So the first condition may be is the patient over or under the age of 70. That's the condition and it's split based on the answer. And the other condition that's important to know when you're running this algorithm is when do you stop growing the decision tree.

So one of the techniques that's very commonly used in decision tree analysis is called recursive binary splitting. So what this does is it basically is an iterative process. There are different splits that are attempted using a cost function. And really what the algorithm is trying to do is basically determine with the different splits how much accuracy is going to be lost in the model. And it chooses the splits that essentially costs the least in terms of accuracy, and that's how the algorithm thinks, and that's how the algorithm works. And the user can actually pre-define what the maximum depth is of the tree, and it can also-- that maximum death basically refers to the distance from the root of the tree to the leaf. So how many conditions are there, how big of a decision tree is it.

There's also techniques that can be done to help improve the performance of the model. So this is called pruning the tree, which is basically you're removing features or branches of the tree that have very low importance. They don't really add to the predictive power of the model. So by removing those you help further optimize the performance. The major advantage I would say of decision trees is they are very easy for clinicians to view them, and the visual interpretation allows you to really make sense of what the algorithm is thinking.

Now just like with other models, there is a there is this problem of over fitting which means if you have an overly complex decision tree, you can be susceptible to the problem of over fitting. So this is an example of a decision tree. So the algorithm will ask, is the patient over the age of 70. If they are over the age of 70 with that condition, essentially it'll move to the next factor that's predictive. So then it asks is the left ventricular ejection fraction less than 30%. Is it 30 to 40 or is it greater than 40%. And then you can see the decision or the leaf of the tree is mortality or survival. So that's what the algorithm is predicting, and based on these various decisions, it will predict what it perceives the outcome will be.

Now if the patient age let's say is less than 70, then it takes you to another condition, which is do they have diabetes mellitus or not. And if they do, it asks about dialysis and so on and so forth. And in this manner, this decision tree's able to ask you several conditions that eventually lead to a predicted outcome.

An other form of supervised algorithm is called neural networks. So this essentially consists of inputs and outputs and there's a series of what are called hidden layers in between. And the hidden layer basically consists of a varying number of neurons, which has connections between the inputs and the first hidden layer. First hidden layer, second hidden layer, and so on and so forth until you have the final hidden layer, and then the outputs that are being predicted. And each of these neurons within each of these layers represents its own individual model, and there's different incoming features into it, and there's different weights that are assigned depending on how predictive that individual model is.

Again, visually looking at this I think helps people understand what this algorithm is showing. So on the left side you basically have three neurons within an input layer. Those are fed into a hidden layer that has a series of five neurons and then the second layer with three. So you can see all of these arrows represent every single interconnection between each of these nodes within each of these layers. And so each of those will basically represents a separate model, and there's different weights that are assigned to each of those models depending on how predictive each of those individual models are.

And eventually you get to a final hidden layer. In this neural network there's two hidden layers just for simplicity, but you could have a series of hundreds of hidden layers. And eventually you come to an output layer. And that's what you're trying to predict. So it's a way, essentially, to-- really it's a way to combine several individual models in a complex manner essentially come up with an output layer that's being predicted by the algorithm.

This is a more popular and more recently developed algorithm that our group has in particular been interested in using for clinical risk modeling. So it's called extreme gradient boosting. This is also a supervised model. So you have inputs and you're trying to predict outputs.

It's what's called an ensemble learning method, which means it aggregates the predictive power of multiple algorithms. And what extreme gradient boosting does is it iteratively builds a stronger model that uses a collection of weaker models that are basically are typically going to be short decision trees. And in each iteration the algorithm will take the difference between the strong model, and the ground truth, and basically it will train a new weak model that predicts those residual values.

And what then happens is it takes these weak models that are predicting the residual that values, and adds them into the strong model. And essentially it'll wait how much that weak model is contributing to the overall model based on the number of outcomes that are mislabeled. So in this way it iteratively builds a really strong model that's using all these weaker and weaker models.

Now when we talk about evaluating machine learning models. So how do we evaluate how well one of these models performs? There's a couple of ways to do that. One way, and this is something that's done and just in general risk modeling as well, is to randomly split the data set that you're working with into a training set, a validation set, and then a testing set.

So the training set-- and this can be done in various proportions. It's sort of user defined. But in general, the training set will have the most number of patients. And you use this and you use this to develop the initial model and run the initial algorithms. The validation set, which will be a smaller proportion, is used to test the model. So once the model's developed then you tested on this cohort of patients and the validation set. You do what's called tuning of the hyper parameters. So this is how the algorithm learns, and you can change a lot of these criteria in the very software that you use.

And then you basically make the model stronger in the validation set. You tune it and you incorporate it, incorporate the changes into an improved model. And then finally you set aside a testing set. In the testing set, this is basically the final evaluation of the model. And so once you have your optimal model that's built from the training and the validation set, then you test it in the testing set, and you can see with a variety of measures how well that model performs.

Another way to evaluate a machine learning model is what's called k fold cross validation. So this is particularly useful if you don't have that many patients in your database. And what it basically does is it takes your data set and it randomly splits it into k folds. K is set by the user. A popular one would be five-fold validation or 10-fold cross validation.

One of those folds is then going to be held out for testing, and the remaining folds are used for training. And this process will repeat a k number of times. So in 10 fold cross validation this process repeats 10 times. And then you basically take the average performance across those 10 folds.

So this diagram helps explain that. This is 10-fold cross validation. You can see each of these blue bars represent the nine folds out of the 10 that are going to be used for the training set, and the orange bar represents the one fold that is held out as a testing set. And this process randomly repeats 10 times, and then what you do is you take the performance from each of those testing sets that was tested 10 times, and you take the average performance and that's what you get for your model evaluation.

So hyper parameter optimization or tuning, if you remember this is done during the validation set phase if that's how you split and evaluate your data. So hyper parameters are essentially factors that can control the learning process of the machine learning algorithm. These can be altered manually by the user, and this is typically in a software program like R or Python, which are popular software programs for running machine learning. Or there are default settings that are provided in these software programs, and those can be adjusted as well by the software.

So some examples of hyper parameters include support vectors and support vector machine, weights and neural networks, or the number of leaves or the depth of a decision tree. In terms of evaluating the performance of the machine learning model-- so a lot of these are very similar metrics that are used when we're evaluating traditional risk modeling, like logistic regression. Really there are two key components that you want to look at when you're looking at the performance of a machine learning model. And these are discriminatory power and calibration.

Discriminatory power refers to the ability of the model to identify patients who have a particular outcome, versus those who do not. And it really answers a question, which specific patients in this study population will have outcome x. That outcome can be mortality, it could be renal failure, it could be stroke, it could be any kind of clinical outcome that you want. And discriminatory power refers to the ability of the model to predict that outcome.

And what this is evaluated with is what's called the area under receiver operating characteristic curve, which has a variety of names that are used in the literature. So AUC, AUROC, or the C index. So the C index is a value that's between 0 and 1. And if the value is 0.5, that essentially will tell you that there's no discriminatory power of the model. It's essentially a coin flip. If it's 0.70, that shows reasonable discriminatory power of the model. When you really start getting into the 0.80 and above range, that's a model that has very strong discriminatory power, and a value of 1 indicates perfect discrimination.

So the figure below actually highlights in AUC or AUROC curve. So on the y-axis, you have sensitivity of the model or what is the true positive rate, and on the x-axis, you have 1 minus the specificity of the model, which is also the false positive rate. And you can see here the dashed red line refers to the line of no discriminatory power, or a C index of 0.5. And you can see the blue line is this particular model, which has an AUC of 0.81, which indicates strong discriminatory power. So the more that blue curve bends to the upper left of this graph, the stronger the model is.

Calibration is another important measure of a machine learning model. So calibration refers to the ability of a model to assign an average risk of an outcome accurately to a population. So really what we're talking about when we look at calibration is the observed to expected outcomes in a model. So a model that's very well calibrated will have observed and expected outcomes that are very similar to each other, and you can actually calculate a ratio of observed to expected outcomes.

The test that's most popularly used to evaluate calibration is called the Hosmer-Lemeshow test, and that test can be automatically run in many of these statistical packages, software packages. And if the p value is non significant, so if the p value is over 0.05, that suggests that there is no significant differences between the observed and expected outcomes, and that suggests good calibration. So you want a non significant Hosmer-Lemeshow test to show a well calibrated model.

So this is an example of a calibration plot. So on the y-axis, you have observed risk and on the x-axis, you have expected risk. The dashed blue line, which is a 45 degree angle line in this plot, refers to perfect calibration where observed risk will equal expected risk at each point. This is an example of an XGBoost or extreme gradient boosting machine learning model where its calibration is depicted with a solid yellow line. The green dots represent each decile of risk and the bars on those green dots represent the 95% confidence intervals. So you can see in this model, there's very good calibration, and that each of those green bars spans across that line of perfect correlation.

Accuracy is another metric that's used to evaluate machine learning models. So accuracy mathematically is defined as the sum of the true positives and the true negatives, divided by the sum of the true positives, true negatives, false positives, and false negatives. So it's really letting you know how well can you identify both positive and negative cases, and this is what's called an accuracy curve. So to develop an accuracy curve you have to develop-- you have to define a threshold for how you're going to then predict the outcome.

So for instance, a threshold of 0.02 would let the model know that, OK if the risk is less than 0.02 as predicted by the model, we're going to predict a negative outcome. If it's over 0.02 we're going to predict that as positive. And so then what you could do is plot this accuracy curve which will let you know on the y-axis what your accuracy of the model is, versus the threshold that you've defined on the x-axis.

Another important metric is what's called the precision recall. So precision is the true positives divided by the sum of the true positives and false positives, and recall, which will be on your x-axis in this plot, is the true positives divided by the true positives plus false negatives. In other words, recall is equal to sensitivity. So this plot basically will plot precision on the y-axis, recall on the x-axis, and it's another measure that's used to evaluate models.

So one thing that's looked at using the precision recall graphs is what's called the F1 score, which is two times the precision times recall, divided by the sum of precision and recall. And basically the optimal F1 score, so the higher the F1 score, the better the model. And the optimal F1 score is calculated at what's called the break even point, which is calculated where you-- this red line, the solid red line that you see, which is a 45 degree angle curve on this plot, where the precision recall plot will intersect with that red line is where the optimal F1 score is.

So here you can see this is a model that's comparing logistic regression model, which is an STS risk model in blue with the machine learning model, which is in yellow. The XG boost model. And you can see when you calculate the F1 score it's improved with the XG boost model. You have a higher precision and recall at that break even point.

So now to the applications of-- that I think provides a pretty general overview of how machine learning algorithms work, how we evaluate the models. And now to really focus on the applications of AI and machine learning in cardiovascular health care. So I really see three major areas that AI and machine learning can be applied, specifically to cardiovascular health care. One of those is automated imaging interpretation. So this is a way for us to develop software that utilizes machine learning that can automatically interpret images that we use in everyday care. These would include things like chest x-rays, EKG, echocardiography, coronary CT, and invasive angiography as well.

So this was a study that looked at chest X-ray. So these authors essentially looked at 14 pathologies that can be detected on chest X-ray. This includes at atelectasis, cardiomegaly, consolidation, edema, effusion, emphysema and so on and so forth. And what they did is they developed an algorithm using deep learning, which is a method of machine learning. And they had a validation set of 420 images, chest X-ray images. And they compared the performance of the machine learning algorithm to nine practicing radiologists.

And you can see here in this graph, these are the AUC curves for the machine learning algorithm, which is highlighted in purple. And then the green lines, which are the radiologists reads. And basically what you see is that when you take a look at the radiologists and the machine learning algorithm, overall in these 14 conditions the algorithm had equivalent performance to board certified radiologists in ten of the pathologies. It was actually better in one, which was at atelectasis. And it was worse in three. And but most importantly-- so overall, I would say the takeaway to that is these machine learning algorithms actually perform pretty well.

But I think the most important part of this study was that the average time to interpret the 420 images was 240 minutes, or four hours for radiologists, and it took the algorithm 1 and 1/2 minute to do all of that. So again, we don't want to sacrifice accuracy when we're doing this, of course, because we're dealing with real patients. But this is a very promising study, and it shows that these algorithms are able to do this very quickly and do it very accurately.

Here's another study looking at EKGs. So again, a deep neural network was used by the authors, and they looked at 12 rhythm classes that in over 90,000 EKGs from over 50,000 patients. And they compared this to cardiologists reading the EKGs. And if you take a look at the algorithms here in the first column, you can see that the C index or the AUC is incredibly predictive from the machine learning algorithm. It's over 0.9 in every single instance, and really reaching 0.99 in some instances, which is incredibly predictive. And then if you take a look at the F1 scores, which are again another metric of performance, you can see that in the sequence and set data sets that there is excellent performance of the algorithm as compared to cardiologists.

This is another application, so coronary CT, the advantages of coronary CT are that it's non-invasive aside from having to give a contrast load. And it helps us get more insight into the level of patient's coronary disease. So in this first column, you see the coronary CT images, the last column are the invasive angiography images. In the middle two the first column-- So panels B, F and J, those are the machine learning based FFR measurements. So we're measuring fractional flow reserve using a machine learning based algorithm. And in C, G and K those are the computational flow dynamic fractional flow reserve estimations.

So you can see here, the machine learning in the first patient, which is represented in the first row, estimates 0.73 for the FFR. That's what the computational flow dynamics estimate, and invasive FFR it's the same thing. 0.74. And you can see across this series of patients that it's similar. Similar measurements and estimations of FFR. So this was the-- for that particular study, this represents the AUC plots for the computational flow dynamics measurement as well as the ML based measurement, as well as visual interpretation of the CTA. So you can see that the worst performing one was visually looking at the CTA, which had an AUC of 0.69. The computational flow dynamic and machine learning models had AUCs of 0.84. So they had very strong discriminatory power.

Now there are some challenges, and this is an area of active research, including my own research, which is how do you use machine learning to interpret video based imaging. So video based imaging would include things such as invasive angiography an echocardiography as well. And there are specific challenges from a machine learning standpoint to developing software that can interpret these things. And that includes non rigid defamation, vessels segment ID, and really as it applies to videos the asynchronous nature of having multiple view videos. We all know that even though we have standard views in both of these imaging modalities there are subtle differences between operators and how we obtain these images. And the majority of studies that have been done thus far in these imaging modalities have really operated on still images from these modalities.

Another broad area of application of Al and ML to cardiovascular health care is in natural language processing. So natural language refers to the ability of computers to read and understand human language. And the real application I think in this subcategory is utilizing machine learning to extract data in an automated fashion from electronic health records for early identification of diagnoses for predicting outcomes, and also for quality control in terms of how well we clinically document.

So this is an example of a study that was done where natural language processing algorithms were running an electronic health record, and the authors used criteria that were developed from the Framingham study for heart failure. And they essentially took a look and they were able to find that at the time of diagnosis of heart failure, if you looked at the electronic health record for the one year preceding that official diagnosis, 85% of patients had some trigger or some meeting of the diagnostic criteria prior to that when you ran the algorithm and you were able to run through their electronic health record. And they developed a pipeline that basically allowed them to extract these specific Framingham criteria, and you could take a look when you look at the precision recall in F score of both affirmed and denied cases of heart failure that we really have very solid performance with this machine learning algorithm.

This is another study that was looking at atrial fibrillation. So again, natural language processing algorithms were developed and run on an electronic health record. And these plots basically show the positive predictive value, negative predictive value, sensitivity and specificity of identifying specific medical conditions that are important in assessing risk in patients with atrial fibrillation. And basically the authors use natural language processing and they calculated the Chad score, which will basically determine the risk of stroke in somebody who has atrial fibrillation, as well as the has blood score, which is a risk of bleeding for patients for patients with atrial fibrillation. So both of these scores are used to determine and anti-coagulation strategy in patients with atrial fibrillation.

And you can see that there is the gold standard, which is marked by the dark gray bars. And that's the score and the percentage of patients that had those scores, versus the algorithm which did this in a very automated fashion. They ran the algorithm in the electronic health record and they calculated these scores, using those conditions that it helped identify and extract from the record. And you can see that there's actually very good correlation between what the algorithm can automatically calculate, and somebody manually going in and calculating these scores.

And here's another study that basically looked at-- so patients, when they self report their symptoms and what's documented in their electronic health record. So this was a very elegant study, and what they showed was if you take a look at the columns that are marked as kappa, the kappa coefficient is basically a marker of correlation. So the closer it is to 1, the more correlated that measure is.

So when you take a look, these are patients self reporting chest pain, dyspnea, or cough. And you take a look at what's documented in the care providers note, there is very weak correlation, which is alarming. If you take a look at the kappa coefficients for instance for chest pain, and they stratified as well into age and sex. But if you take a look, the kappa is 0.52 for dyspnea 0.46 for cough 0.38 that's very, very weak correlation. So that's an alarming thing in terms of quality control for how we document things in our electronic record.

Now they develop a natural language processing algorithm and they ran it, and you can see that the kappa coefficient was much improved. In terms of being able to automatically annotate some of these things and what patients are reporting, so you can see that for chest pain it improved to 0.78, for dyspnea and improved to 0.74 for and for cough to 0.69.

The third area that I think AI and ML will be important in cardiovascular health care is in predictive analytics. So predictive analytics refers to risk models that we develop, and it could be for any outcome. It could be for operations, it could be for procedures done in the Cath lab, it could be for various outcomes such as mortality or complications. It could be models for readmission to the hospital. These models are essential to the field of cardiovascular medicine both from a cardiology standpoint as well as from a cardiac surgery standpoint. They have significant utility in program evaluation, so if you're able to develop models that better estimate risk, you can have a better measure of what you're observed to expected outcomes should be. Quality improvement therapy selection are also based on these models.

So here's an example of the use of Al and ML in predictive analytics, so this was a risk model that was developed by the authors to look at the risk of major bleeding after PCI. They used five-fold cross validation, and basically the existing simplified risk score had a C index of 0.77. And you can see that they ran a variety of models and a blended variety of models and really they were able to get an improvement to 0.82 with a blended model that used gradient descent boosting. And when you looked at their calibration plot, which is shown in the bottom right, they did have improved calibration utilizing the machine learning model.

Here's a study looking at the risk of acute kidney injury after PCI. So again, a variety of machine learning models were applied to this data set. When they looked at logistic regression the C index was 0.717. That improved with machine learning all the way up to 0.759, And you can see this dashed line in the middle. Anything to the right of that is better than the baseline model. So you can see a variety of these machine learning algorithms had stronger predictive capability as compared to the baseline model.

And here the model that they felt performed the best was XG Boost, which is extreme gradient boosting. And it shows that this is their calibration plot, and it shows that it's a very well calibrated model. And the slope, the closer the slope is to 1 the better calibrated the models. You can see that the calibration is actually improved with the machine learning model.

And this was interesting plot that they had. So model one, which represents their baseline logistic regression model, and model eight which represents their XG boost or machine learning model. So they divided the patients and stratified them into what the predicted risk was based on the logistic regression model and based on the machine learning model. And then in the middle there are the actual rates that were observed in the patients.

So you can see, if you go down that left side and model a predicted risk, so when it predict-- when the machine learning model predicted less than 5% risk, and the logistic regression predicted less than five, you get 2.8% which is an agreement. And if you take a look at that row for the model eight that predicts less than 5%, most of those patients are below 5%. If you look at the machine learning model for 5 to 10, and you look across that row, most of the patients fall between 5 and 10, regardless of what logistic regression is predicting. And so on and so forth.

Now if you do the opposite, which is you look at the columns. So you look at model 1. When model 1 is predicting less than 5%, you see that there is a variety of what's going on with the patients based on what the machine learning model is predicting. So overall what this graph is showing is that there is improved predictive capability in terms of observed to expected outcomes with the machine learning model.

The shaded bar-- the shaded cells in this plot referred to when there was congruency between the logistic regression model and the machine learning model. And you can see, so when it's less than 5% in both, the risk is 2.8. When it's 5% to 10% in both risk is 7.1. When it's 10 to 25 risk is 16.5, and so on and so forth. It falls exactly in line. And so this is an interesting concept because machine learning may not be used just to replace existing models, but can also be used in a supplementary way like this when, you can see if there's congruency between models, and that gives you stronger affirmation that you're predicted outcome is accurate.

This was a study looking at in hospital mortality after cardiac surgery. So the euro score is a well established score that's used for risk modeling in Europe for patients undergoing open heart surgery, and it utilizes logistic regression modeling. And you can see in this data set when they were predicting in hospital mortality the euro score had a C index in the low point sevens, and the author used a variety of machine learning algorithms and they showed that with an ensemble of machine learning algorithms they were able to improve the C index all the way to 0.795.

And this is looking at the risk of acute kidney injury after cardiac surgery. So again, a variety of machine learning models were run and compared to logistic regression modeling. And essentially you can see that the AUC was most improved with EX boost, or with gradient boosting where the C index was 0.78. And these were the flat of the AUC curves for that study. And you can see again the more you are closer to that upper left corner of this plot, the more predictive the model is. And you can see that red plot, which refers to the gradient boosting machine learning algorithm has the most predictive capability.

This is a study that we actually performed here at UPMC where we looked at over 11,000 patients that were undergoing open heart surgery, and we compared the STA, which is the Society of Thoracic Surgeons, they have risk models that are well-established and have been used for decades for looking at programs and for therapy selection and so forth. The STA risk model in predicting operative mortality in our data set was 0.795, and this was improved utilizing a machine learning model XG Boost to 0.808. and this panel below shows the calibration plot which again demonstrates that the XG boost model had very well calibration when we developed it.

The other neat thing about XG boost in particular as an algorithm is it gives you insight into what it's thinking and what it's using to make its predictions. So here this is a plot basically showing the relative importance of each individual variable to the predictive model in the XG Boost model. So you can see in this particular analysis that we did looking at operative mortality that the strongest predictor was the serum creatinine. Other important factors included weight, age, ejection fraction, intra-aortic balloon pump use, peripheral arterial disease, and other co-morbidities, as well as presentation of disease. A lot of these risk factors we would identify as being important, and many of these are included in the STS risk models as well.

This was an interesting plot in my mind, because this is essentially looking at what was the correlation between the STS predicted risk for each individual patient, and their machine learning predicted risk. So you can see that the correlation coefficient only shows really moderate correlation of 0.65. And perhaps what's even more interesting is if you specifically look at the patients that actually died and had operative mortality, there is actually a pretty weak correlation between what XG boost had predicted, and what STS it predicted. And that correlation coefficient is 0.47.

So if you take a look, and let's say you were to set a binary threshold of 15% you were going to predict that a patient had an outcome. So in a patient-- in patients that had less than 15% predicted risk from STS, but had over 15% from XG boost, you would have picked up an additional 51 patients that would have been accurately predicted. If you do the flip side, you would have only picked up 26 patients that were low predicted risk by XG boost, but high by STS.

So in summary, AI and machine learning in health care is expected to grow exponentially, not just in the field of cardiovascular health care, but really in all medical fields. As it relates to cardiovascular health care, the potential applications include automated imaging interpretation, natural language processing and data extraction and quality control, as well as predictive analytics.

There are challenges to implementing AI and ML in health. One issue is public trust. Are patients going to trust machines to make health care decisions for them? Some of the machine learning algorithms are truly a black box. You have no idea how they're predicting things, they just predict them. There's also an issue of privacy. Is there going to be disclosure of sensitive personal health information, or is the algorithm going to-- for patients who don't disclose that information, is it going to make some prediction based on them not disclosing it or automatically assume that are one way or another?

There may also be biases in treatment recommendations. For instance, if you have a machine learning algorithm that's a black box, and you don't know how it's predicting, but let's say it identifies race or gender as a predictive risk factor, it may bias the treatment recommendation without the public knowing why it's biasing it. Because that is a risk factor for an outcome. We also need to make sure we have continual evaluation to make sure that the decisions that are made by AL and ML algorithms are not harmful to patients. And finally there is a lot of legal loopholes that will likely need to be jumped through to get software to clinical practice. And the FDA is actually considering Al and ML software to be considered as medical devices, and go through that route.

So in conclusion an AI and ML in health care is in an exciting phase. This is really a blank canvas. There is a lot of active research going into this, and the current role of Al and ML in health care is being shaped as we speak. Research in this field is really growing exponentially, but the application of research to clinical practice will require careful thought and input from multiple stakeholders.