(PLEASE NOTE: Transcribed automatically by Vimeo, mistakes are possible/likely. Our apologies.)

TRANSCRIPT - GR 10 14 22 *"Sepsis and Machine Learning"* – Andrew Barros, MD from the University of Virginia

UVA IMR

00:12:55All right, everyone. We're going to go ahead and get started

- 00:13:00uh welcome to Medicine Grand rounds today. Beautiful fall day out there uh today. I'm. Excited to present Dr. Andrew Barrows from the division of pulmonary and critical care. Medicine.
- 00:13:10Dr. Barr is well known to Uva. Having completed his undergraduate Medical School Residency and Fellowship. All here
- 00:13:17along the way he's found a little bit of extra time to also complete a master's degree and start a family as well.
- 00:13:24Dr. Barrows is highly regarded amongst our Residency program as one of our best educators, and he has been instrumental in teaching a lot of us in the room. Critical care medicine, both on the words as a fellow now in attending as well as in the classroom through our acute care lecture series,
- 00:13:39his research focuses on a data-driven approach to prevent harm in the intensive care unit by merging his clinical expertise technical training and computer sciences and the electronic health Record.
- 00:13:51An effort to promote this important work. Dr. Barrows recently received a Career Development Grant through the Integrated Translational Health Research Institute of Uva, otherwise known as I thrive.
- 00:14:01Dr. Barr is a wonderful combination of intellect, curiosity, and humility. He's a bright future ahead of him, and we're excited as a metal community that we can all be a part of it. Please join me, giving a warm welcome to Dr. Barracks.
- 00:14:18I think everyone comes up and says that was like a fantastic introduction to me, and I feel very honored by it. But I feel
- 00:14:24flattered. Um, I'm going to start by.
- 00:14:33There we go, so I just want to acknowledge. You know, some folks who have really helped support me in my career, development and presentation of this talk in my research um
- 00:14:41sort of two broad objectives for what we're going to talk about today. So you know we're going to talk about machine learning and sepsis. But I have to come across with concrete things that I hope you'd be able to do uh identify the end of this talk. Some appropriate metrics for how we evaluate
- 00:14:54machine learning and sepsis, and then describe some of the barriers between our current state and sort of widespread adoption.
- 00:15:00Um.
- 00:15:02As a junior faculty member. I think I'm contractually obligated to say I have no disclosures, but I'm open to obtaining some um, and then, you know there's It's interesting. So you know one of the things that the C in the office is to say is whether we're going to talk about off level uses of medicines and devices, and this, software as a medical device where exactly these machine learning fall is sort of the Wild West. So I'm not going to intentionally discussing off label Use
- 00:15:25Um, that being said that on the twenty-seventh of September of this year, the Fda sort of really some new guidance of how they are interpreting statute about how software should be uh a framework

for regulation of software as a medical device. Um, and it's really complex. And a ton of stuff might be considered a medical device. Now and then it would have labels and off label uses um. But this is also really new, so nothing applies here.

- 00:15:49So the sort of scheme for the talk. It's going to be four broad parts. Um, a little bit about Why, Sepsis, and why Sepsis is, I think, an interesting problem to try and apply this to.
- 00:15:58We'll spend a little bit of time talking about machine learning. Not certainly to get into, You know the nitty, gritty details, but enough to sort of understand what some of these terms are that are being used in this, like numerous publications that come out with this now
- 00:16:10and then sort of why machine learning specifically for sepsis and does it work uh? And then a section about why isn't it everywhere? Why isn't every hospital using a machine learning model for their sepsis screening.
- 00:16:23So I think everyone here is sort of familiar with sepsis, and it's frequency and it's complications. But I think it's worth emphasizing a couple of points. Um.
- 00:16:31So using the sepsis three definition, we'll talk about some of the definitions later, but are more the most recent definition of sepsis. Um. It's President, about six percent of adult hospitalizations. And across all those hospitalizations about fifteen percent of people with the diagnosis of this will die during that hospitalization.
- 00:16:48Um, This figure on the right comes from a Medicare cohort uh looking at basically different uh different points in time with with Covid nineteen to the reception. So, for example,
- 00:16:58make this thing work right. That point right. There represents everybody in this Medicare cohort who is diagnosed with Sepsis and January fifteenth, and you can see that there's pretty impressive mortality for some of these subsets. Right? So um
- 00:17:10for the in the Medicare population the one week mortality was close to forty for folks with septic shock, and also surprisingly, this mortality risk persists beyond after their admission. So, you know, three years after their admission. Eighty. The Medicare beneficiaries, who had septic shock have died.
- 00:17:26Um! And so you know, I think there's an incredible burden of morbidity uh with sepsis.
- 00:17:34We also think of sepsis as being time sensitive. Uh on this slide on the left side of the slide I've just quoted directly from the guidelines. Um, the surviving set to skylines about sort of antibiotic treatment, and our time to treatment, and in severe sepsis and septic um, you know, they say, uh
- 00:17:51administer, and my current meals immediately. I deal with the one hour of recognition for both sepsis and septic shock. It's listed as a strong recommendation with very low quality of evidence.
- 00:18:01Um! On the right side is the classic figure from the Kumar paper that looked at time from Hep, attention to administration of antibiotics and showing this sort of linear time dependent relationship. The laundry, the longer the gap was between when someone developed shock, and when they got appropriately treated for their separate shocks. More likely they were to die.
- 00:18:19Um! I think it's sort of worth when we think about this being explicit about? What are the steps that lead from
- 00:18:26development of a condition to the actual treatment of it? Right that they have to develop sepsis needs to be identified. We need to order treatment. It's got to be administered. And so there's a lot of steps along this pathway,
- 00:18:36and you know, while there are some
- 00:18:40small sort of observational trials of these things, but nobody has randomized people to a delay in antibiotics and septic shock like That's not a trial. We'll ever be able to run. We're never going to know definitively about that time-sensitive nature.
- 00:18:52Um! And you know you would wonder, too. Is this all about the Is this all in antibiotic administration? Or is it some component of recognition,
- 00:19:01you know? Is there just something different about these folks that have a different sub phenotype of sepsis that they are harder to recognize. It takes us longer. Um!

- 00:19:10So um! Another sort of example of this. So, in two thousand and thirteen New York State passed a law requiring all our hospitals to develop a sepsis bundle, implement it and report metrics, and they had some minimum required bundle elements. But they did not say you have to use this bundle specifically,
- 00:19:28and so uh, after some time had passed, and they could report or analyze all the data. This was looking at sort of the end product of time to bundle implementation and the likelihood of an inpatient mortality, and so on the bottom. We have um sort of odds ratios per hour,
- 00:19:45and you want to be on the left side of the line. But it looks, you know most of these points are on the right. And so you know another sort of example of the longer it took the hospital to implement that full bundle which was
- 00:19:55um measuring a lactate, drawing blood, cultures, and antibiotics was in their one-hour bundle the more likely you were to die,
- 00:20:03but intriguingly, if you look in sort of one of the supplements.
- 00:20:07They also, for sensitivity, split it out by the components of the bundle
- 00:20:11right and so on the left side we have antibiotic administration. The same figure. What's the time to the Association for the Risk of death for a one hour delay?
- 00:20:21But on the right side is for measurement of lactate, and I I find this interesting, because, as far as I know, lactate has no therapeutic value. Right? It's only diagnostic. And so you know, I think one of the things is
- 00:20:33is that a reasonable market to infer, when Sepsis was identified that the mayor's lactate itself was measured. Um! Both of these are associated with an increased risk of death and delay.
- 00:20:42And so
- 00:20:44there's lots of things that come into antibiotics. But maybe identification is an area where we can make some progress.
- 00:20:53Um! This is the nineteen ninety-two sepsis one definition, the original. So sort of pre nineteen ninety-two um.
- 00:21:01Everyone told me what back to Romeo was, and we had an idea of what septic shock was. But we did not have a widespread operational definition for sepsis, and the goal of these authors was to produce a definition that could be used to coordinate clinical trials and conduct additional research. And they they're very clear here that their goal was to build a a very sensitive but not super specific
- 00:21:21definition for sepsis,
- 00:21:25and throughout time our definition is changed. Right so on the left side we have, you know the first one thousand nine hundred and ninety-two definition, which was essentially just suspected. Infection plus inflammation.
- 00:21:35And Now to the two thousand and three definition, it's
- 00:21:39suspected infection plus inflammation or organ failure to the two thousand and sixteen definition which is really infection plus organ failure alone.
- 00:21:47Um. And I think one of the challenges is the rationale for these definitions right? So none of these were trying to make a set of diagnostic criteria. The one thousand nine hundred and ninety-two. They were trying to make a definition. They would help them facilitate clinical trials in two thousand and sixteen. It was trying to make a definition of a population that was really high risk for bad outcomes,
- 00:22:07and to facilitate sort of benchmarking and coordination between hospitals.
- 00:22:11Um. And so, you know, we'll talk a little bit about this later, but
- 00:22:18what definition we choose sort of might change who we want to predict right? So the sepsis three cohort of bad infection plus organ failure is a population that's highly likely to die from their sepsis
- 00:22:31that's a subset of people who probably need antibiotics right? There's lots of people have infections who do not get to the organ dysfunction stage. And so if we try and screen or diagnose, or only identify sepsis three, we recognize. There's a ton of people who
- 00:22:44we're not going to identify there.

- 00:22:48Um, Also, as we go through a couple of these different models, we're going to note that as we've gone through time, definitions of change, and it's, I'm not convinced that that sepsis three totally supersede sepsis two,
- 00:22:59and you'll see this temporal trend of how these models go.
- 00:23:04I think it's also reasonable just to reflect on what our Sepsis management has done over the last thirty years, right? So in one thousand nine hundred and ninety-two. Thirty years ago we had antibiotics, fluids, and as oppressors,
- 00:23:13um and now we have antibiotics and basic pressures,
- 00:23:17bundled care and looked atable and ventilation.
- 00:23:20Um! And so you know, the nih has spent
- 00:23:23three nhs a ton of money. We've run a ton of clinical trials. We've had a ton of agents that failed. We had one agent that hit the market and was withdrawn. But we have like really no sepsis specific therapeutics. And So you wonder if we're really going to move the needle on Sepsis mortality that if identification might be one of these areas where we could improve performance.
- 00:23:46So. Um, we've sort of that under way. Let's pivot just slightly. We'll talk a little bit about machine learning. We have sort of a nice foundation that we can discuss the art as it is now.
- 00:23:56And so you know, just some vocabulary components. Um, Because we like to use confusing terms so often understand about features. So features an input to our model. Um, Almost anything can be a feature. But most of the models require numeric values, And so there's a bunch of steps what commonly called pre-processing to turn
- 00:24:16whatever it is that we're actually observing in the real world, and to a numeric input that we can feed into our model, and those pre-processing steps are important, and are, as you know. Well, there's sort of the before the modeling, and the step are required. If you want to be able to do this anywhere else,
- 00:24:33the other thing is, we have our target or the thing we want to predict, and it could be a continuous, variable right like we could try and predict somebody's height. It could be a sort of binary categorical. Do they have pneumonia? Do they not? It could be a time to event.
- 00:24:47Um. But regardless of what the event is, we need some way of ascertaining this, which some people call the ground truth, or their gold standard. For some of these things. It's not that complex right. If we're doing height,
- 00:24:56measuring tapes exist, we can just measure people. We have a direct measurement there. But you know, for example, in Sepsis, I think the labels are a little less clear uh, and that's one of the things we have to think about here.
- 00:25:09You know, the other thing is most of these models output probabilities. So twenty-five percent chance that this person has pneumonia. We need some way of turning that into a Yes, no, we're going to do something. We're not going to do something. And so there's a lot of implementation process that comes after these things.
- 00:25:28What is a model? Um!
- 00:25:32I thought about the sentence for a while, so I think it's. It's an artifact of a process we use, make predictions, and I use artifact in a very specific current sense of not artificial, not like artifact on the Kg. But rather sort of this object created by a human process that we at the end then do something with,
- 00:25:50and because, you know, one way we could think about. This is uh our machine learning process. Is this machine where we put data in one side we turn a crank.
- 00:25:58On the outside other side comes a model that we can then implement or do something with, and you can rerun that machine many times, and you can come up with different models. But you know this, this, this artifact itself is what we give to people to go do things.

- 00:26:11And that could be a nomogram, an equation, a decision tree, a set of weights in a neural network. What it is, I think doesn't matter as much as it's. Just the process. The thing that comes at the end of this process.
- 00:26:24Um, I take a deliberately broad view here, because I think a lot of things end up being models, and we can share evaluation strategies among them.
- 00:26:34So in machine learning, particularly for sepsis, you end up seeing sort of two broad categories of of models that are used. Um. One is random forests which will find a little bit talking about. So um! This is a
- 00:26:49training data set about some penguins. And so we can imagine a hypothetical example here, where we're trying to identify what species of penguin we have a gentoo or not, chattoo penguin,
- 00:26:59and I think you know naturally it makes sense as we would look at that, we could pretty reason quickly, sort of derive a set of decision rules in our head For what's Gen. Two versus not Gen. Two, perhaps, you would say, Look at mass and say, Have your penguins, mass greater than four thousand four hundred grams or ten to the other ones, are not.
- 00:27:17And so you know, when we use that in a computer to technique, we call that a decision tree. Um. And you know, I think that one of the downsides to trees is, you can imagine, if we haven't a tree of unbounded length with
- 00:27:31many, many thousands of decision points, we could very quickly learn very detailed patterns in the data that did not generalize. Well, that didn't help us predict for the next patient. They really just helped us predict in that set that we used.
- 00:27:45So the random a forest approach is to take decision. Trees and sort of do two tricks, maybe three tricks. Um! One is to just sort of randomly get rid of some of the features. So if we were looking at this, we would make a different tree, if weight wasn't there as a column
- 00:28:02um, and another one is to sort of randomly get rid of some of the rows, and so you can avoid certain like high leverage to steal six term rows or rows that have a lot of importance. If, when you randomly drop and drop those by chance, um that helps improve the likelihood that you'll make some trees. It'll fit outside the data you're looking at.
- 00:28:21Well, there's a lot of randomness in here, and so we should expect that sometimes these trees are not going to work that Well,
- 00:28:26sometimes they're gonna be pretty good. And so the third trick that random for us is, let's just do it a lot. So if we make a thousand trees an average over all the results, what we call an ensemble,
- 00:28:35we can have a a performance overall. That's reasonable.
- 00:28:40This gets us a couple of features that we don't always get. So you know one is that when we think about the break points. If we had a decision point for mass, and then a decision point for a flipper length. What we really done is create an interaction between flipper, length, and mass, right so that
- 00:28:56you know you could have presumably even different outcomes for the same flipper length based off mass there.
- 00:29:02Um and I wasn't going to get into the method that much. But because these things are fit sort of algorithmically taking what's the best option at that particular choice from the data it's given. Um! They can find these interactions without you having to explicitly say, I want you to consider mass and flip the length.
- 00:29:20Um, you know, this is like a a a recurring theme. Oftentimes machine learning that we build these ensembles, or relatively weak learners that do well together.
- 00:29:30Um! And there's a lot of models that use tree-based methods like this.
- 00:29:37Another one is deep neural networks, which we're not going to spend a ton of time on here, but essentially for deeperural networks. We have a bunch of inputs.

- 00:29:45What I end up thinking of, and oftentimes are just a bunch of linguistic progression models in the middle, and we can add as many of these hidden layers as we want. And so the output of one becomes the input of the next. So we have, you know, if you think about each one of these edges as being
- 00:29:58a parameter, a knob, something you have to tune in the model. All of a sudden we have thousands of parameters, and we have this very flexible model that can talk about interactions between any number of features at the downside of it, not always being super clear that what comes out is related to what went in.
- 00:30:16So, as we said, all these models come out with probabilities. But we want some way of talking about. How do they work? What you know? What's their evaluation? What their metrics? Um. And I think we naturally fall into this mindset. If we think a lot about diagnostic testing right, do they have a murmur, do they not?
- 00:30:31Um. And that we have sensitivity, specificity. But those are really at a threshold, right? So, we could arbitrarily choose a threshold like, say, fifty. It's not necessarily the best one for our scenario, But you know we have these continuous predictions about risk that are coming out of these models, and we need to turn them into binary um binary things. We can act on.
- 00:30:51And so um, you can talk about just like we do about diagnostic testing sensitivity, specificity, positive predictive value, negative predictive value, positive likelihood ratios, negative likelihood ratios.
- 00:31:04Uh, but we have to recognize that it's at a specific threshold, and how you choose. That threshold is not really related to the model. That's an implementation decision,
- 00:31:14and so it's hard to know what the best threshold is, and we'll talk a little about that later. I think most of us care, most probably about positive predictive value, but we recognize that has a lot of things baked into it that are hard to control. What's the threshold? What's the prevalence of the components?
- 00:31:32So
- 00:31:33you know, Oftentimes we ask ourselves two big questions about this models and how they work. So how well does it classify what people sometimes call discrimination.
- 00:31:41Um. Two cases reliably have scores or probabilities that are better than non-cases,
- 00:31:48but also equally accurate, is how accurate are the probabilities or calibration. So if I took a group of people with ten percent risk to ten percent of them have the outcome
- 00:31:56Um, Because, particularly with these rare events. If we're not going to be making super, you know, if we're going to predict
- 00:32:04a risk of ten percent which might be two orders of magnitude higher than the baseline risk. Um! We should recognize that a lot of those people with ten percent risk are not going to have the outcome.
- 00:32:14They only have ten percent chance
- 00:32:17sometimes to sort of bring in these merge. These two things together. You oftentimes see figures like this: These curves uh area under the curve.
- 00:32:26The pedantic person to me wants to say you. Then you say, what curve into this area under the receiver, operating characteristic curve from
- 00:32:34cold radar research. Um,
- 00:32:37And so uh, just to review it. So, this is actually some data behind. Uh, we built a
- 00:32:43a machine learning model that looked at white blood cell parameters. So the machines we use in the lab right now to actually measure white blood cell counts. Are these really complex purpose built flow cytometers, and they use some of the data to result it to the chart. And some of the data just stays on the machine and includes
- 00:33:01all these interesting channels about
- 00:33:03how angry are the nutrients to make the pathologist angry that I use that term. But how much Dna replication is happening in it which ends up being the

- 00:33:10fluorescent side scatter of the neutrophil gated channels and support that. But so we tried to build a classifier for predicting present on the mission Sepsis based off the first Cbc. With differential that was collected in the emergency department,
- 00:33:23and that's what we get here.
- 00:33:24And so um, just to sort of navigate everybody through these figures. Right on the left we have the receiver operating characteristic. So on the bottom is the false, positive rate. Right? So um, and then on the y axis becomes our true positive rate,
- 00:33:41and we sort of. I like to start when I think about it by thinking about the two extremes, right? So that top right corner, the hundred percent false, positive rate, One hundred percent identification of cases
- 00:33:51uh is what happens if we set our threshold to basically zero, right? If we say everybody has the condition, we have a boat with a false, positive.
- 00:33:59But we correctly identify everyone in the data. Set
- 00:34:04the bottom left corner right is the really high threshold. If we set the threshold that's one hundred and one no one can ever get up there we will never have a false positive. We're never going to call case.
- 00:34:16We also never identify any of the true cases,
- 00:34:19and then, so now we could consider every possible value between those two points, and figure out the false, positive rate, the true positive rate
- 00:34:27where the sensitivity and specificity and plot a point.
- 00:34:30Um, if we built a really dumb classifier and just flipped a coin every time we had our classifier output had no relation to our condition, you end up with a straight line between these two components. Right? So our probabilities end up, equally distributed. It doesn't matter where you pick your threshold
- 00:34:47uh you end up with a straight line, and what we want, the ideal classifier would be as close as possible to this top left corner, right would have
- 00:34:55zero percent false positive rate, and would find all the cases. And so
- 00:35:02we plot the continuum of points along here, and then we can visually look at them. There are statistical methods for saying one curve is better than another, but oftentimes this falls down sort of a visual look, or how many decimal points do you end up needing trying to decide that A is better than being.
- 00:35:18And so for this model we tried sort of three different sets of parameters, the base model is just age, and the total white blood cell counts.
- 00:35:27The reduced is age, the total by itself, count, and all the stuff that comes on the regular differential,
- 00:35:32and the full is adding in all the weird research parameters that we don't currently use,
- 00:35:36and it improves classification, ability, right. We get closer to the top left corner.
- 00:35:40Excuse me,
- 00:35:42um. On the right we have what's called a precision. Recall curve, which
- 00:35:46you technically can also have an area under that curve, which is why I say, we shouldn't talk about a You see it should be a
- 00:35:52um, and so the precision recall curve at the bottom. We essentially have what they're calling recall or the percentage of cases correctly identified or positive cases identified.
- 00:36:02Um
- 00:36:04and then on the Y-axis becomes essentially the positive predictive value.
- 00:36:09And so this
- 00:36:13this point down here
- 00:36:15basically just determined by the prevalence in the in the um, the population you're studying right. So if you set your threshold relatively low, you will identify one hundred percent of the cases. But

your process of predictive value, and this set ends up being about ten percent, which is what it looks to, which is what

- 00:36:33you what when our baseline prevalence is. And then, as we sort of change our threshold to have less sensitivity. We're going to increase our the positive predictive value.
- 00:36:44So in this particular model, if you come down to, if you say Well,
- 00:36:47i'm happy, only identifying forty of the cases,
- 00:36:51and then it, you know, for the
- 00:36:53full, modeling it up with a positive predictive value of about twenty five percent,
- 00:36:57which is isn't bad, you know It's a several full increase over the baseline. It's not great, but it's there um! And so in on. Both of these in the on the legend and parentheses it's showing you the first actually sorry on the left side. It's the air end of the curve on the right side. It's just the average precision across uh all those values.
- 00:37:17An alternative formulation of the area under the receiver operating curve is what's called the c. Statistic, they mean measure the same thing. It's just like a different way of thinking about it. And so to see statistics is, what's the probability that if you took a random case and a random non-case, that the case is going to have a higher predicted probability in the non-case,
- 00:37:37I think both of these sort of emphasize
- 00:37:39the problem with relatively rare events, that if
- 00:37:43you know your event rate is less than one, you can get a pretty good classifier by just saying no one ever has the condition you can end up with. A pretty reasonable roc.
- 00:37:54Um, although your positive value won't be that great,
- 00:37:59That being said, the majority of literature in this world, just for that just presents a you see, for the uh roc curve,
- 00:38:07I think the other thing is um sepsis screens typically not single shot or deal. It's not. You show up to Emergency Department. We run our fancy model on you. We make a prediction. We evaluate how we did. Oftentimes we are making multiple predictions on the same subject throughout an entire hospitalization,
- 00:38:23and I try to be polite here that I think reasonable people disagree about the right way to characterize that scenario.
- 00:38:30So this is a table from uh a validation of the epic sepsis model. We'll talk a little bit more about that model later. But I’m just trying to emphasize here that we can see as we change our time horizon. Are we looking for ever over the whole hospitalization in the next twenty-four hours? Next twelve hours
- 00:38:49we change our income out outcome incidents right so thankfully, you know less than zero point. Two of the of the people in this data set we're developing sepsis and the next four hour window compared to six percent for the whole hospitalization which fits with. You know what we saw on that sort of previous work.
- 00:39:09So but it, you know this is an example right as the preval as the um prevalence goes down, we see that actually for the same model, the A, you see, looks like it comes up,
- 00:39:19but the positive predictive value has fallen, and this is the problem with rare events. Um, that I think we really care about positive predictive value.
- 00:39:27But people always want to tell you what their au.
- 00:39:32So
- 00:39:34why machine learning for sepsis?
- 00:39:37Um,
- 00:39:39I think we're all relatively familiar with news. This is news, too, because it's easiest to find a nice figure of news to the version two up there than it was the original one. Um! This is part of what I’m going to call family of like clipboard models that you could imagine that these are things you could score with a piece of paper and a pencil. There's no fancy techniques behind it.

- 00:39:57You look at each category, each row, which is a parameter assigned points based on where they fall some of their score. You have some action threshold score greater than eight, for example, and you do something with it. Um,
- 00:40:10you know we look at this and say, Well, there's seven rows, and up to three points, so your maximum score would be twenty one if you were terrible in all this categories, or
- 00:40:18because it's actually nineteen. Um
- 00:40:21at
- 00:40:23So There's a finite number of predictions that we're making on this populations that probably does not have the full flexibility to describe all of human disease and nineteen different predictions.
- 00:40:35But how does news do for sepsis? Uh! Well, that's not bad, despite it. Not really being a Sepsis model, right? It's just a general decompensation in the acute care model. Um! It actually has a reasonable a You see that we just talked about how you see, like really not the right metric for this it's also the easiest one to find. Often
- 00:40:55this is sort of an interesting different formulation. This is from that same paper from before, from Lou, and so on the x-axis we have sensitivity. So Um! What percentage of our cases are we identifying? And then on the Why, access? They just said, what proportion of the patients in their data set would have cost that threshold.
- 00:41:12Um.
- 00:41:14And so, you know, I ideally. Um. We want to have
- 00:41:18really sensitive, but not have a ton of identification of their cases. And so,
- 00:41:23when they show it here great the gonna be as far as possible from this line. Um unsurprisingly, Sir doesn't do that great. But news does okay for identifying or to the compensation. And there's been some observational studies looking at things like um out of the out of the Icu d conversation, showing that it helps there.
- 00:41:44But
- 00:41:45but why machine learning for this?
- 00:41:47So you know, one thing we come down to is this is
- 00:41:51you're gonna notice a recurring theme here. Is that because a lot of this work was done here, there's a lot of papers for Dr. Mormon on them. But this from Travis Moss. Those are the groups looking at a code of folks, and just how frequently are we recording observations on these patients, because this is one of the things that are we really in a big data problem, or are we in a just a definitional problem.
- 00:42:09So for this um eight thousand emission data set, they ended up being that there was about um one point, eight million labs, four million vital signs. And then, if you use things like continuous cardi, respiratory dynamics, or continuous monitoring, and up with about thirty-five million features observations. When you use fifteen minute features,
- 00:42:29and you know, two point: six billion heart beats and six hundred million breaths. And then the other thing is, just look at the features here, right the sort of time the um rather than sorry the distribution between observations. So
- 00:42:44for just vital signs, you know, typically we're not obtaining vital signs more frequently than every hour, every four hours or every six hours. And so there's just a lot of time between when you have new data to make predictions, and a lot can happen in six hours.
- 00:42:59By the same token, Well, labs seem very informative about underlying disease states. We just don't check them that much. There's a limit to how much blood you can send to the lab. We check labs daily in most scenarios, and you see that distribution there,
- 00:43:13but breathing and heart rate, and these continuous features are recorded all the time, and so that provides this opportunity to use more data to help make better decisions.
- 00:43:23From here. You know

- 00:43:26this becomes so. What if we're going to step aside for second, and look at not adults? So this is an original description of heart rate characteristics in very low birth weight neonates with sepsis,
- 00:43:40and
- 00:43:42uh
- 00:43:44uh, the A. B and C is as the same infant on three different occasions. Um some time before Sepsis. I believe it was like six hours before, and then just before the diagnosis there. And
- 00:43:56I like to figure because you can kind of visually see. It looks like the right Dynamics are different from there.
- 00:44:00Um! And this got turned into a product called Hero. That um
- 00:44:05brings all these components together
- 00:44:07uses harming dynamics to predict their risk of sepsis, and then just displays to their user where their risk of sepsis is compared to the baseline.
- 00:44:15And you know, I think the thing about heart rate dynamics in this scenario is that are you doing with a very unique set of physiology that
- 00:44:23those hearty variability are oftentimes
- 00:44:27I mean, I don't treat kids, but described as like essentially um, partly accelerations like fetal decelerations, because very low birthday infants share a lot in common with fetuses, even though they're no longer in the room.
- 00:44:38Um!
- 00:44:39And so there was a they conducted a trial. This this is the uh, as far as I can tell, the only randomized trial of Sepsis analytics that has ever shown a mortality benefit.
- 00:44:50So uh, it went to nine. I see use. They just threw the display up on the wall randomized beds to having display, or not having to display and randomize about three thousand babies, and have a mortality benefit
- 00:45:02um absolute risk reduction of about two percent.
- 00:45:05And then the majority of that, all that was essentially in the sets subgroup.
- 00:45:09Um.
- 00:45:11These are kind of this is kind of interesting, because
- 00:45:15you know, to steal a quote from some other folks. Um,
- 00:45:17the models don't prevent sepsis h hand washing does. And so this is just showing early detection and trying to accelerate the treatment timeline.
- 00:45:27Not unsurprisingly. People have tried this in adults, in fact, so many, so that we have a Meta analysis. Um,
- 00:45:34this is the these are all the randomized trials. These are of a non random.
- 00:45:40It's like really off. I like point way over there. These are all the non randomized trials, and I mean I'll comment on a couple of things. One is that you know the majority of them trend on the left side of the null. But you know there's basically no positive randomized trials.
- 00:45:54Um. The other thing is that the three thousand infant and Hero is basically more than all of these trials combined.
- 00:46:01Um. And you know, in in that I also see here, when you take all these trials together, There's a suggestion that automated sepsis alerting, using The majority is using machine learning techniques and produces mortality um by about thirty, which is not insignificant.
- 00:46:18We're gonna look at this one a little closer just because it's the largest of the sets um.
- 00:46:26And so
- 00:46:28this model was called uh Insight, and they and some of any design that they You look at that. You say you Csf has seventeen million inpatient counters across five years. But no, they actually use a lot of encounters putting emergency department and filtered out of time to come down to create this model to basically predict sepsis to

- 00:46:45um subsist to your sepsis and septic shock, based off nothing but vital signs.
- 00:46:49And just for fun they use a tree model. And
- 00:46:52you know again, having said our. You know the Rcs on everything. Um! These were sort of how they were able to classify folks, and it seems to do reasonably well and better than others, other models for identifying things like septic shock ahead of time.
- 00:47:11And then they did a non randomized study where they implemented this across eight hospitals, and it's primarily before after uh, and showed a reduction in mortality length of stay and read mission rates, which is kind of impressive.
- 00:47:24Um,
- 00:47:26Another sort of non-randomized trial again. Sort of done here. Uh is so. This is comment, so comment uses, uh vital slabs and continuous monitoring integrates them all together and produces a um sort of an ensemble of models to talk about the risk of respiratory decomposition cardiovascular instability.
- 00:47:45And then it's presented on uh with a with a trend. So uh this trend becomes that over time they've moved from here to here. They've gotten bigger and sort of showing increasing risk in the subject,
- 00:47:59and it was implemented in the city with again A very similar Put it on the wall that people look at it. No specific alerting strategy.
- 00:48:07Um, and then they use the medical icu for comparison across the same time.
- 00:48:11Um! And you did. They showed a reduction in septic shock sort of one of this hypothesis that if we're identifying people earlier, you know again, the model is not going to prevent sepsis, but perhaps early identification, more prompt treatment might prevent septic shock or these bad complications of sepsis.
- 00:48:30A little plug for what we're doing here in the Uva process is that we have this automated service called ramp. Real time analytics and monitoring platform that lets us build our own models, score them on patients in real time and then provide user displays. Um. So this is a sample of what's running for news, plus, which is a locally modified version of news
- 00:48:50um, and showing risk for patients over time.
- 00:48:53Um, as you mentioned that the uh comments being studied, most of you have seen this on the fourth floor in a randomized clinical trial. Um! And hopefully those results will be soon. I Here we just finished a moment.
- 00:49:04Um and payers are starting to, or certainly a mandate sepsis screening. So we will see more of these models in the future.
- 00:49:13So so what are the barriers? So why, Don't? Why, doesn't every institution have one of these state of the art
- 00:49:20automated one at models performing screening right now.
- 00:49:24Um,
- 00:49:26I think one problem is that Sepsis is a difficult outcome.
- 00:49:29I've sort of harped on this a couple of points. But I’m going to spend moment thinking that explicitly. So you know, Steps to three, identifies patients at high risk of death. It doesn't tell us who needs antibiotics, and that sort of
- 00:49:41dissonance between what we want the models to do and what the models are actually predicting means that you know affects how users perceive them and actually implement them.
- 00:49:50You know, I think the other challenges that the signature of organis functions not. We need to sepsis. We struggle with this, when we all know these clinical scenarios about, does this person have a spec infection? Um. But as we evaluate these models, right bad pancreatitis and bad sepsis are might have will have a similar physiologic signature.
- 00:50:07Yeah, one of those we consider a true positive one of those we consider not.

- 00:50:11Um, and related to that. There's just substantial heterogeneity and the cause of people sepsis the response to this, and how they get better. Uh,
- 00:50:21this is a paper from the group here. Trust Moss, where they, you know, fit a whole bunch of models for various conditions in one cohort, and then tested them in another.
- 00:50:30So um on the vertical axis.
- 00:50:34Try to make sure I got this right
- 00:50:36on the vertical axis that the model from where it was fit, and then on the horizontal access. We've checked it in that in that population. So, for example, um!
- 00:50:45This is a neonatal Icu integration model.
- 00:50:49It does not do well at predicting hemorrhage in the Mickey. That's good, right like we would not expect there to be a ton of crossover between the signatures.
- 00:50:58A couple of things here, you know, one part is that it's interesting that for the Neonates
- 00:51:02really, there's a lot chair there like the innovation model seems to do reasonably well for predicting transfusion and sepsis,
- 00:51:10and then in the adults synapses in the mic you
- 00:51:15that's very poorly a predict or sorry if this is sepsis in the sick, you does very poorly at predicting sepsis in the medical icu,
- 00:51:22and the reverse is also true. A model trained on sepsis in the medical icu does not predict sepsis in the search twice you,
- 00:51:30which is, I think, an interesting finding, and I I don't know exactly what right it's not like the bacterial pathogens are,
- 00:51:38you know, Staff Aureus is the same on the fifth floor and the third floor. So this may represent the disease conditions, and you know which set of folks have a post operative inflammation versus something else. Um, and you know, I think is is an important thing to consider as we try and take a model from some other scenario and apply it to where we want to be here.

UVA IMR

00:51:59This is from A. There's a an annual challenge called Computing and cardiology, and in two thousand and nineteen they gave teams a data set and said, Build a machine learning model to predict sepsis, and they gave the teams test Set A and B.

- 00:52:14So the teams were allowed to see both those ones and then set C. They held back. So you the way that competition worked. You submitted your code, your whole thing, they said. Hey, we'll run it. We're going to score it on our data sets. And then they brought in new data, and also scored it on there.
- 00:52:29And so
- 00:52:31the interesting thing here. So each one of these dots represents one entry. So they had, like almost ninety entries overall, and
- 00:52:39this is the best performing model on the full test step
- 00:52:44when they split it up by institution.
- 00:52:47The full test set model here was like number two on test set a and number two on tests at B.
- 00:52:54But on the on the hidden test it all of a sudden fell down way down here to the twentieth,
- 00:52:59which is also sort of a very interesting thing. It's uh observation that no one has a great explanation; for, because, as far as we know, the features of the like, you know the distribution of blood pressure and the thing all the same, across all these data sets um, and you know the other interesting thing is to me at least, is,
- 00:53:15we have a whole bunch of these folks who did not do very great down here on the two. You know tests that they saw

- 00:53:22but generalize really well, they actually did the best when taken to the other data points.
- 00:53:28So you know, I think this is one of the challenges we need good definitions. A good population that we're looking at, and a consistent disease state because these models are probably not super generalizable between institutional institution, disease, disease, unit, and unit.
- 00:53:47Another problem is that people don't always find it useful. So uh pen built this random for sepsis model. The stats are there. It's not a bad mob like a positive predictive value of thirty. For sepsis, I think, is is reasonable, and people would appreciate that.
- 00:54:04Um! They had this interesting thing where they um At the time of the alerts they called the rapid response. They disseminated a survey, then forty-eight hours that disseminated another survey, and they just asked people like, Did this affect your expectation that the patient was going to develop critical illness
- 00:54:20did it result New findings did it change your management,
- 00:54:24and overwhelmingly People said they didn't change their expectation didn't change the management. They didn't find them useful across positions and ours is everybody.
- 00:54:34This is from Trues, which was a um
- 00:54:38as a septic sepsis model. That um has gotten a lot of press recently, and I I bring up this figure. So here the this is the hours before the first organ failure,
- 00:54:50for example, so that the purple ones were identified by both the model and regular practice.
- 00:54:55The yellows were done by just the model. The greens were done just for the practice, and then the Grays nobody found. And I think the one thing here is that there's substantial um identification of of these conditions by the people taking care of them.
- 00:55:10And that sort of brings me to my next phones here, which is like we don't really know what our baseline is,
- 00:55:16So I think what we really care about for automated screening is that incremental improvement in identification?
- 00:55:23But we don't have a great way of telling me like I I can't. The computer can't read people's minds, I can't tell you. When did same Oliver actually diagnose sepsis? Right? We can infer when they order blood cultures or antibiotics, or lactate like it's a pretty good predictor for suspicion of Sepsis.
- 00:55:40But there's some that are like these relatively weak predictors that we just don't. We can't, you know, order to Cbc. Could mean that you're worried. They're bleeding or you're worried. They have an infection. And because we don't have a great way of inferring. What are they actually thinking we're left with just saying they're totally blindfolded.
- 00:55:57And so
- 00:55:59what ends up happening is, I think, when it feels like a party trick when it's this black box, making predictions that sometimes tell you what you already know, and sometimes disagree with you. Users lose confidence in the predictions. And so you know the end of the day is like the models of three patients themselves. So we need to figure out how to
- 00:56:16make these predictions make them useful, and display them in a way that people can synthesize it into their practice pattern.
- 00:56:24I think the other component of this is
- 00:56:29the models can be good, but they're not. They're not going to raise the standard where, I think often has people want their performance to be so. Again, This is the stats on the pen model again. So a positive, likely ratio of thirteen really not bad, right? The reasonable possible predictive value there.
- 00:56:42Um!
- 00:56:44It's a sidebar. It's like surprisingly hard to find good likely ratios for exam maneuvers, but based on some old Jama review. Um exposed bone, and a diabetic foot wound has about a likely ratio of about nine for Auster. So you know, acknowledging the whole baseline risk thing.
- 00:57:02Some of these models have a positive value that, for that's about the same.

- 00:57:06I think one challenge is, we all want performance. That's like a Ct. Scan, right like the like. The ratio for Mri and diagnosing Ostia is like infinite um. But we're getting performance. It's like physical examiners. I I think we can get better. I think we can go from there, but some of the current state models are still, you know, Don't, have the same performance that we want.
- 00:57:28The The final problem is that implementations are really hard. Um. So this is the epic systems model. This is the version one of the model, because after this paper came out they basically took it back and made another version. Um, but this was a This is a commercially available model. If you hand up a compile of cache, they will hand you this model back.
- 00:57:45Um, and they. So this health system side is externally validated. The you know. The reported performance was pretty good. Their observed performance was not as good
- 00:57:54um, and there's a couple of things to sort of for clear. So one. But we already know that you know that make you sick. You probably need to train at least refine in your data set,
- 00:58:04but you know they evaluated for this epic score of greater than five, because that's the threshold that they had already turned on in their institution.
- 00:58:13Um. So at that score, in an eight hour time horizon, we had a number needed to evaluate a seventy-three,
- 00:58:19so seventy-three met calls to find one person in sepsis um, which sort of and the positive predictive value there was relatively for um, spider and reasonable a you see.
- 00:58:31So
- 00:58:32you know, even after we build these models, even after we train them, we still decide what's the right threshold to use? What are we going to do with them,
- 00:58:39and and I think that's one of these gaps between the model buildings, the fun part. All the work is what you do after that. So you know, the optimal threshold finds enough events, and doesn't have too many false positives. But
- 00:58:51enough and too many are relatively subjective, right? And I think one of the things we have to say acknowledged is that
- 00:58:59our mathematical techniques tend to call a correct positive prediction, like identifying sepsis and a correct negative prediction, saying, they don't have sepsis as equal of importance, and we call, you know, we call a false, positive, and false negative is equal. This utility
- 00:59:16um I don't place a lot of weight. My fire alarm not going off when there's not a fire, and I think the perspectives of how would you wait? Those things probably vary across administrators, nurses, physicians, patients. And this is one of these challenges of it's not.
- 00:59:31We often talk about. This optimal decision point is mathematically closest to that top left corner. But there's a lot of implementation behind it, and that's one of the barriers to this widespread screening.
- 00:59:41So my conclusions, So you know, I think machine learning is definitely a problem for Sepsis detection. There's a ton of data There's time patterns, machine learning will find them.
- 00:59:52We need to better define the outcomes that we care about. Um because general decompensations. One thing Sepsis is one thing,
- 01:00:00but you know our users are going to evaluate the models based off the outcomes that they care about. And we need to make sure we're predicting those things.
- 01:00:08Um sepsis a hard problem. But I think early detection is the way we're going to improve outcomes.
- 01:00:13Um. The data you generate is more important than the models, and the implementation is more important than either.
- 01:00:20I think a good model would be targeted to a population in the illness. So a reasonable scope that you could evaluate, uh, understandable. The people at the other end know why the scores being generated and the things that input to it makes sense. To us, physiologically,

- 01:00:37um and ideally, it would mean that you could generalize it in some sense, so that septic shock and the Mickey here would be the same as cfic shock. And m you at Vcu, for example.
- 01:00:48Um, and I think we need to do more randomized trials. Um! When we look at you know we're going back to them. Analysis: a ton of not randomized series. We go back and look at industries that have adopted machine learning, Google, Netflix, for example, Google. Famously. They call them a B test, but they randomized trials
- 01:01:05like, get an rct of sixty shades of blue for one of their interfaces, like I don't know that every randomized trial has to be nih funded with tens of thousands of people. But we should have less pre-post founding better implementation results.
- 01:01:21And with that,
- 01:01:23thanks for your attention, happy taking questions

Unknown Speaker

01:01:33chat to

UVA IMR

01:01:44thanks for great talk. Andrew Um.

- 01:01:47One of the things that I think we struggle with is the definitions of Sepsis are pretty poor, and generally created by like a Delphi process of people sort of deciding what is important to them and what they're trying to predict. Do you think there's a role even before early detection of Sepsis, of trying to refine
- 01:02:05the definition of sepsis, using something more like a machine learning model or a
- 01:02:10something else. I don't know much about
- 01:02:24a bunch of biomarkers in sepsis, and created two sub-enotypes a hyperinflammatory and a hyper inflammatory phenotype that boiled down to
- 01:02:31by my butcher if I just try and come up with the top of my head but three features and intriguingly, you know, in her data, when they went retroactively and looked at sales, which was a trial of Statins and A. Ds.
- 01:02:43The group who they would their model would assign to the high inflammatory phenotype,
- 01:02:47responded to sympathy. Satin and the low, inflammatory group did not now admittedly like hip hops generating a retrospective. But I think we all recognize that there's some types of sepsis, and then machine learning has this promise here. Um, There's another great paper that
- 01:03:02I also wanted to thought about, including it. Sort of split people up into a couple of subgen the types of sepsis based off clinical parameters. And
- 01:03:10yeah, So I I do think there's a way something like that, just like we talk about in a specific disease. Well, I think maybe all substance is not the same. And you know I don't want to make light of a pandemic that cause millions of deaths. But I think we recognize right that, like Covid, you know most, a substantial fraction of people's covid qualified for sepsis that that disease is different than
- 01:03:30Stephanie. It's back to Remia. Um. But yet, from a diagnostic label we call them the same.
- 01:03:41I think that the models the difference you saw in Sepsis and the Mickey versus the sick. You could have been related to the some phenotypes of sepsis, and
- 01:03:51it's It's probably a combination of some phenotypes. I wonder also, too, about practice pattern elements, right? And so I think most folks building models nowadays recognize that, for example, Um, you can't use people ordering a lactate as a reasonable marker for that freight, because clearly ordering

lactate is a marker. Bounce your sepsis. I suspect there are more subtle differences between the two units that might indicate a clinical suspicion for Sepsis, And that's one of the challenges

- 01:04:23things that have uh stuck out to me um with the data that you showed from Pen with the thirty percent positive predicted value sepsis, and about seventy percent of people saying it did not impact their uh their management.
- 01:04:41Do you think it would be do? Is there significant overlap uh in those two groups? Um, meaning like, do you think people were correctly saying, this has I've identified the thirty percent of people who
- 01:04:55have Sepsis on my own, and so the model didn't add to it. Or do you think there were a fair number of patients who did have sepsis? But the clinician did not uh correctly use the information from the model.
- 01:05:08I I think it's a combination of, and not to go back to that particular one, but to the true picture here, Right? So all the purple ones were model and clinician agreement. Right? Both the model identified sepsis and the people treating them. Um and the yellows were just identified by the model
- 01:05:27Early on um. But the greens were just an inender routine screening, and I think the one challenge is when you look at the Ac. For the model and totally identifying steps. It looks really good. But what we really care about is this incremental performance. Um.
- 01:05:41And and you know, just like people rapidly
- 01:05:45conditioning the wrong word, but rapidly grow to this behavior of like overriding like Bpas, like the Osa Bpa. I think a lot of false, positive sepsis alerts for people who already think you have make you lose confidence in
- 01:05:58that minority that the model is identifying, that you haven't found yet.
- 01:06:03Um, so it's all about trying to make the right prediction to the right person at the right time in an actionable way.
- 01:06:11That's all of clinical decisions for it.