Dr. James P. Rathmell: Hello. I'm Jim Rathmell, Professor of Anesthesiology at Harvard Medical School and one of the Executive Editors for Anesthesiology. You're listening to an Anesthesiology podcast that we've designed for physicians and scientists interested in the research that appears in the journal.

Today we are going to talk with the senior author of an original research article that appears in the October 2018 issue. With us today is Dr. Maxime Cannesson. Dr. Cannesson is Professor of Anesthesiology and Vice Chair for Perioperative Medicine in the Department of Anesthesiology at the University of California, Los Angeles in Los Angeles, California.

Dr. Cannesson is the senior author on an article that appears in the October 2018 issue of the journal titled "Development and Validation of a Deep Neural Network Model for Prediction of Postoperative In-hospital Mortality." Dr. Cannesson, thank you for joining us.

Dr. Maxime Cannesson: Thank you for having me, Dr. Rathmell.

Dr. James P. Rathmell: Well, congratulations on the publication of your work and for conducting one of the first, if not the first study, that employed deep neural network-based machine learning, a form of artificial intelligence or AI that we’re hearing about every day nowadays. And you’ve used this to predict in-hospital mortality after surgery.

Now, I want to ask you to start by giving listeners just a brief primer on artificial intelligence and specifically this form of AI using neural network-based machine learning. Can you briefly explain how a machine “learns” using this approach?

Dr. Maxime Cannesson: Yes, I can give you a brief explanation about how in general a machine learns. They actually learn the same way we human beings learn. We learn how to recognize patterns; for example, as a physician, if you are a very experienced physician as compared to a trainee, it usually takes less time for the experienced physician to recognize when the patient is sick, even going to the preop area and looking at the patient before we bring the patient for surgery, when you’re an experienced physician sometimes you are able to detect very briefly how the patient is going to do during the surgery.

That’s what we call clinician, the difference between what we have in the patient’s chart versus what we see when we look at the patient. How do we learn these as physicians? By seeing a lot of patients. So, it means creating a database and seeing the outcome of the patient during or after the surgery.

And in our brain, which we learn when we train, is to associate the patterns that we detect in some patients, so in our database with the outcome that we are willing to improve for the patient.

So, computers do not do things much differently. In order to develop machine-learning algorithm, you need big databases of patients or of subjects and you need to define an outcome that you want to predict and then the machine learns the same way we do by analyzing a lot of patients and associating features, variables in the data sets and detecting associations between variables that are going to be related to the outcome that you are trying to predict.

So, in essence, machine learning are trained to reproduce what we do as human beings; it means learning from massive databases and associating patterns and association of variables with an outcome of interest.

Dr. James P. Rathmell: So, you used a machine-learning technique into the study we’re going to talk about today and we’ll get into a little bit of the nitty gritty. But first, what was the research question you set out to answer? What was your hypothesis?

Dr. Maxime Cannesson: So, in this specific study the questions was pretty simple because as you know, this is increased in health care and especially anesthesiology, that’s a pretty new field, right? So, we want to start with outcomes that we can easily identify and probably one of the easiest outcomes to identify is in-hospital mortality. Why? Because it’s easier to track patients’ mortality during the hospital stay than after their hospital stay. And second, because death is a pretty easy-to-define outcome.

So, in this specific study our goal was to develop a deep neural network algorithm that would predict at the end of the surgery, and that’s something we can discuss further, but we used the data from the patient at the end of the surgery and this algorithm was trained to predict in-hospital mortality.

It’s critical, really, to have outcomes that you can identify very easy. For example, if you want to predict postoperative myocardial infarction, that becomes much more tricky because the definition of myocardial infarction can change from one center to the other or from one paper to the other and the way we measure myocardial infarction after surgery is based on troponin, for example, and we don’t have troponin on every patient.

So, for this specific paper the idea was to predict in-hospital mortality, the reason being because it’s very simple to identify. And so, a meaningful outcome, of course.

Dr. James P. Rathmell: Perfect. So, in-hospital mortality. But your hypothesis was that machine learning could do as good a job as more complex models that we’ve had previously. Is that what you set out to test?

Dr. Maxime Cannesson: So, I think in our mind, really, the idea was to see whether it is feasible to use machine-only algorithm and deep neural network to predict in-hospital postoperative mortality.

And what we wanted to do, so, is if we do that we have to compare it to risk classification tools that are already existing. So, first they are divided in two groups: you have prop risk stratification tools to predict postoperative mortality or intraop. The prop scores that exist are as simple as the ASA score which is actually not very simple, but that’s a pattern recognition score that clinicians use to predict mortality.

We wanted also to compare deep neural network to what we thought would be more classical approaches to predict mortality like logistic regression techniques and we wanted to compare it to also administrative risk prediction scores like the RSI and RQL. And we also wanted to compare it to intraoperative risk prediction scores and especially in our study we used a surgical Apgar score.

Dr. James P. Rathmell: Alright. I want to spend some time talking about exactly how you conducted this study, your methods. Walk us through the data set that you used and how you defined the endpoint for the model. I think you’ve done a good job already, but we’ll go through it again. And then the features of the input to the model.

Dr. Maxime Cannesson: So, the database that we are using in this study, one thing to understand—and it’s the limitation of the study—that we use only a single database to train, develop and validate the
model. We used a database that we’ve developed at UCLA, especially at the Ronald Reagan Medical Center which is the big hospital within the UCLA System.

We studied about 50,000 patients; we excluded patients who went for outpatient surgery and for all of the patients we had in the database we only took the first surgical event; we excluded further surgeries in the same patient. And at the end on the database that contains about 150,000 patients, we selected about 50,000 of them.

The endpoint was to predict in-hospital mortality. We defined in-hospital mortality as death in the hospital. The data source we used to defined death was the electronic medical record. And that seems trivial but depending on the data source you would use to define an outcome as simple as death, you may have different incidences of the outcome that you’re trying to measure.

So, in this study specifically we defined in-hospital death as death in the hospital with the data source being the electronic medical record. And then what we did with that, we selected a set of 87 features that we would use in the deep neural network model to predict in-hospital death.

These 87 features, mostly intraop features, that were selected by a group of expert clinicians who thought that these features would be more likely to be associated with an outcome like death after surgery which shows you further to some extent this could be seen as a limitation because you have a human being, expert clinician, selecting the features that they think are going to be associated with death.

That’s – in other words, you could call that a supervised machine-learning model where the user selects the input features for the model as opposed to an unsupervised model where the machine would randomly select any kind of data available to predict the outcome that you want to predict. So, this is a supervised model.

And then what we did is that we used these 87 features, we initially focused especially on the intraop features, so at the end of the surgery we take all these features and we feed them into the model and we predict mortality. And what we did then is that we took a preoperative score—and that’s the ASA score that was scored by the physicians—and we enriched the model with the ASA score.

The idea is really to reproduce what clinicians do is that before surgery you predict an outcome using the ASA score or any other score; but things happen during surgery that changes the risk of your patient. And so the goal was to then putting in an intraoperative score on that would adjust the patient preoperative score.

Dr. James P. Rathmell: Okay, we’ll come back to that in a minute what really predicts and how you did the enrichment, but I want to talk about these data sets. You’ve got 50,000 or 60,000 patients, they have 87 different features, features like age, blood pressure, heart rate, laboratory values and there’s missing data or there’s data that’s just not plausible. You talk in your paper about a systolic blood pressure of 400; it’s just impossible. So, how did you handle missing and implausible data?

Dr. Maxime Cannesson: That’s a very interesting question because how do you handle artifacts, how do you handle what we believe misinformation or data that sounded credible, right? And that’s – although it’s a big question about what is a credible data, what is an artifact and what is a phenomenon, whatever.

But what we did was pretty simple. For some of the (sounds like: intervariables), we defined what we as clinicians believed would be a not credible value like as you mentioned a systolic (sounds like: captured) pressure of 400. Anything above 400 would be eliminated from the model and the same way we would have extremely low value for all other variables like temperature or whatever.

The selection of the data was, in our study, actually pretty simple. That’s not the most sophisticated part; that’s something that’s based on clinical expertise where we set boundaries defining what we believe is a credible value.

Dr. James P. Rathmell: So, you select credible values and then you put it at the extreme of what would be possible. So, now you have a clean data set; I’ve put “clean” in quotes because you’ve cleaned it up by substituting plausible for implausible and average values for missing values.

But the incidence of mortality is under 1%; it’s a very, very rare event that you’re looking for. So, that poses a challenge because then the machine has to learn to differentiate these rare cases from the rest of the patients who survive until discharge from the hospital.

Why does this low incidence of the feature you’re trying to predict present a challenge for machine-learning approach and how do you handle it?

Dr. Maxime Cannesson: In this study that was one of the biggest challenges, right, that you’re trying to really detect or predict an outcome that happens extremely rarely and the risk of this is to have what we call misclassification. And the lower the incidence of the outcome you’re trying to predict, I would say the smallest or the most precise the outcome you’re trying to predict is, the more likely you are to misclassify patients.

So, in this specific study, you have multiple ways to go around this. So, the first one is going to be the level of training, the time it takes to train the software. The idea that when you develop a deep neural network, the deep neural network during the training phase is going to output the statistics. And what the system does is that it compares the statistics that is output by the system to the actual statistics in the population. And the training is going to be focused on teaching the model how to get closer and closer to the real actual statistical value of the outcome you’re trying to predict with the output of the model.

When it’s very low like this, the training takes much longer than when you’re trying to predict an outcome that happens in 40% or 50% of your patients. Why? Because the precision of your prediction has to be extremely good. Basically in this study the low incidence increased the duration of the training.

The second thing is that you have statistical ways to artificially increase the incidence of the outcome, in this case the mortality. So, you would select (sounds like: separate) prediction of your data sets that have a higher incidence of mortality and then what you do is that you start training the model on this specific population.

And then step-by-step you expand your data set to then achieve the low incidence of mortality that you have initially.

Dr. James P. Rathmell: Alright, enough of the nitty gritty. Can you tell us how you actually divided the data to train the machine and then how you tested the resulting algorithm? What did you compare the machine learning algorithm to output against? You hinted to that earlier, some other models that you used. But talk us through.

Dr. Maxime Cannesson: So, developing and training machine-learning models to predict an outcome is – the methodology is actually pretty simple. So, the first thing is that in an ideal world you need two databases with two different patient populations: the first database is a database that you’re going to use for the training, development and internal validation of your model.

So, in our study we actually only had one database which is a UCLA database. The way we did and that’s what’s done in many, many studies on this topic is that you divide the database in two different subsets: one is the training and development subset, usually that’s about 80% of your database. And the other one is an internal validation database that’s usually about 20% of your database; that’s exactly the proportions we use in our study.
How do you separate the development and training database and the internal validation database? You usually randomize. In this case, you randomize patients; sometimes in administrative databases they randomized hospitals or even countries or states or whatever. In our study we randomized patients to training and development to an internal validation; it was 80% versus 20%.

Then what you do, you use the 80% of your patients’ population, the subset, for the training and development. When you’re close to getting your model trained, then what you do, you validate it internally in your internal database and that’s what we did in this paper. And that looks nice, but one can argue that testing an algorithm that you developed on one part of the population and testing it on the other part of the same population, it’s not very surprising if it works well, right?

And, unfortunately, in our study that’s what we have; we only tested it in an internal validation database. But the state-of-the-art machine running of prediction statistics of science say that you need an external database with a different population and validate your model on the statistical validation. And that’s really when you can validate what you’ve developed.

In our paper we did not do this extended validation. If you look at another paper that’s not using like deep neural network, but that’s using logistic regression, that was published in Anesthesiology a few years ago focusing on the score that’s called the POSMOP score that was developed in France on millions and millions of patients, theirs, though, only did the internal validation without external validation.

And what people have to understand is that to make these scores really actionable and really trustworthy, they have to be tested externally.

Dr. James P. Rathmell: Alright. So, you hinted at this. We already have some pretty sophisticated logistic regression models that can fairly accurately predict in-hospital mortality and from these same or very similar data sets. What makes machine learning different than these logistic regression models?

Dr. Maxime Cannesson: The first thing to understand is that actually logistic regression is one form of machine learning and if you had to compare logistic regression to a deep neural network, the difference is what’s recorded in a number of layers in your network. The logistic regression is a neural network with only one layer of neurons; deep neural network is a machine-learning method, like logistic regression with multiple layers of neurons.

So, what it means, and I’m trying to make it simple, is that when you develop this model, you have (inaudible) about 87 features and the machine is made to associate randomly all these features and these combinatorial features are going to be used to predict in output statistics.

In the deep neural networks, these combinatorial features are going to be recombined with other recombined features at different layers of neurons. So, it’s almost an infinite ability to train your model. In logistic regression, you have only one layer of combinatorial features. So, it’s potentially very, very different because the deep neural network has much more power to try to predict the outcome.

In our study we wanted to compare the deep neural network to the logistic regression model and see if the deep neural network would do much better. And what we found is that using the same futures to predict in-hospital mortality, the deep neural network actually doesn’t do much better than the logistic regression.

People may wonder, then, why do we bother to use these kind of techniques? My personal conclusion is that the value of these machine-learning algorithms explodes when the complexity of the data sets becomes more and more complex.

And EHR data sets usually can be big in volume but are not usually very, very complex. For example, one way to complexify your data sets is to add time series: taking the same values, but every minute, so I reset again, so I read two minutes, whatever, so to augment the complexity of the data set.

In this specific setting, I will hypothesize really that deep neural network will do much better than logistic regression, but we did not test this in our paper.

Dr. James P. Rathmell: Okay. So, more to come. But you also spent some time whittling the predictive muddle down from 87 features for each patient to far fewer. How many data elements are needed to accurately predict mortality? And can you tell us some of the most important features that predicted mortality?

Dr. Maxime Cannesson: This concept relates to the concept of data parsimony. The idea is that when you develop the model, we tend at the beginning to take the larger features to train the model, but the goal is to go down to the minimal set of features that we can use with the same prediction to predict the outcome we want to predict. That’s called data parsimony. You want to decrease the data set.

Why is it important? First, because it makes usually the system a little bit more accurate and precise, you have less misclassification. The problem is that if you get too parsimonious, you don’t have enough data, enough features, and then you have a misclassification that also gets higher.

In our model, specifically, what was pretty clear and it’s very important, is that one of the features that predicted mortality the best was actually the ASA score. And that seems very surprising to some because papers have criticized the ability of ASA to predict outcome. In our study, actually, the ASA is doing a pretty good job.

And then you have other variables, like age, hemoglobin at baseline before the beginning of the surgery is a pretty strong predictor, the use of vasopressor, low mean arterial pressure. So, a few of the features of a heavy weight in the deep neural network. But the ASA before surgery is probably the heaviest one. So, it means human beings do a good job at predicting the outcome of their patients.

The way I like to explain is that the ASAs also a neural network, but the metrics for the neural network is biological, it’s the brain of the physician. What we are trained to do with computers is to replicate this pattern recognition but using a digital matrix.

Dr. James P. Rathmell: Well, that’s reproducible too; faulty computers amongst humans, you know.

Dr. Maxime Cannesson: Yes.

Dr. James P. Rathmell: So, tell me, what did you conclude from your work? Should we all download your algorithm—I mean, I see that you’ve made it available online—and perhaps start calculating the probability that our patients will die before hospital discharge at the end of the surgery? And what might we do with that information? How would we use it?

Dr. Maxime Cannesson: So, first I would not recommend to use the algorithm right now; clinically I would recommend some scientists to test the external validation of this algorithm on the external databases and that’s why we wanted to make this algorithm public so people can use it to validate in their own institution. So, I would recommend clinicians to not use it right now; it’s probably a little bit too early.

The second thing is that let’s take the hypothesis that this algorithm is externally validated and has a very good accuracy. What do we do? We use this data. The first thing is that how do we implement these scores? And what I foresee that at some point the question will be, how do you implement the score in the EHR?

So, at some point the implementation science, meaning the ability of implementing this kind of complex software on your EHR, is going
to be the next stop. And that's not easy because first now we use commercially available EHR, the electronic health record, that are not easy to modify. And second, we usually don't have live data that would generate a score at the end of the surgery yet.

Then the question is, once it's implemented and you have it at the bedside, what do you do with the score? With the prediction of mortality, probably one of the first things that this kind of score are going to impact is the resource utilization is, do you send your patient to the PACU? Do you send your patient to the ICU? Do you send your patient to the floor? Do you discharge your patient home?

I think on this big outcome like mortality, really the goal is to impact resource allocation. But in the future what we want to develop is the risk prediction scores that are going to be very specific to a disease like the ability to predict postoperative acute kidney injury or the ability to predict postoperative pneumonia.

And once you have these prediction scores, the next step will be to develop what we call prescriptive analytics. It means that the software then based on the features that are in the model will be able to detect what features has the heaviest role in the prediction of the outcome.

And this feature is a feature that you can reverse. Maybe the software we propose to act on this feature, not to avoid the outcome you want to avoid. In other words, if we knew at the end of the surgery that the patient has a risk of AKI and one of the features in the model to predict AKI is because this patient did not receive enough fluid during surgery, the system could, for example, suggest to give more fluid to the patient. That's the concept of prescriptive analytics.

And computers may help do that, actually; identify what can be done to reverse the probability that the outcome is going to happen.

**Dr. James P. Rathmell:** It's pretty exciting. So, what comes next for you and your research group? And I want to ask you to expound a little bit. Are there ways that you're employing artificial intelligence in the realm of the anesthesiologist already? And then how do you see AI impacting the field of anesthesiology in the immediate future, the next year or two?

**Dr. Maxime Cannesson:** So, on the research we are doing right now, the main topic is to expand the complexity on data sets that we are using to really enhance the analytic power of this machine-learning algorithm. So, my belief is that using purely EHR data is not going to be able to leverage the power of machine learning. So, using time series, for example, to complexify data sets is one of our research topics today. The other one is to use more complex information and complex information that we add to our data sets are the physiological waveform. So, the waveform that we get from the monitor, and we record them on every patient and then we merge the waveforms’ data with the EHR data and we apply machine learning to waveforms to try to get develop intraop predictive analytics that are physiologically based.

Another arm to complexify the data is to use genomics data and some of the patients we have at UCLA, we have - there is a big project within UCLA that's called the AtLaS project where we connect genomics data before surgery and the plan in the future is to incorporate genomics data in these models. Once we have this complex data set, then we can focus on very meaningful clinical outcome.

In the next two years, what's going to be the implication of machine learning in our clinical life, I think really the first thing, the low-hanging fruit, will be the risk stratification before surgery; that's going to lead to, I believe, resource allocation. Historical data in health care have always been used first by administrator and for resource allocation.

Very little use of machine learning or analytics have been developed initially in a patient-centered way which is sad but it's usually the resource allocation and the administrative user data is easier to issue. So, I think in the next two years the low-hanging fruit is going to be risk prediction and resource allocation.

The next step—that's probably in the next five, ten years—is going to be able to use this kind algorithm to make any kind of provider able to predict the risk for each patient. What I mean by that is I think today experienced clinicians with 30 years of experience are pretty good at predicting the outcome of their patients. The challenge is to have trainees predict with the same accuracy.

And I believe that machine learning is not going to do much better than the experienced clinician, but in the hands of a trainee they are going to help trainees or younger physicians or less-experienced physicians to predict outcome of their patients with the same accuracy of the – as an experienced clinician.

And then I think the long-term view will be to merge predictive analytics, artificial intelligence software together with some form of automation, closed-loop system, and help clinicians work under a more supervised role where the technology will be used extensively to assist, for example, mid-level providers and under supervision of physicians and anesthesiologists.

**Dr. James P. Rathmell:** Outstanding. So, I have a lot of promise as a chair and an administrator to use AI for predictive analytics and resource utilization maybe even tomorrow. But that explanation you give about making nonexperts predict a lot closer to what experts can predict we're seeing across all sorts of different uses of AI. It takes all of those things that take many, many years for experts to develop and makes them available to the novice.

**Dr. Maxime Cannesson:** Yes. But I think that machine learning is not different than any other technology. I use the example of the pulse oximeter. Before the pulse oximeter was invented, it would take years for a trainee to be skilled at diagnosing cyanosis in a patient under the (sounds like: drape) during surgery. The trainee would have to see many, many patients and it would take months and months for him or her to be skilled at this.

And when the pulse ox was invented, you just had to tell the trainee if the number on the screen is there's a 90 and the waveform is appropriate, your patient has cyanosis and then in five minutes you would train someone.

**Dr. James P. Rathmell:** Terrific. I hope today's discussion will lead many of you listening to read this new article and the accompanying editorial view that appear in the October 2018 issue of Anesthesiology. You can learn more about artificial intelligence, machine learning and how it might impact anesthesiologists in the very near future. Dr. Cannesson, thank you for joining me today and for the terrific explanations about artificial intelligence and its implications for practicing anesthesiologists.

**Dr. Maxime Cannesson:** Thank you very much for having me today. Thanks a lot.

**Host:** You've been listening to the Anesthesiology journal podcast, the official peer-reviewed journal of the American Society of Anesthesiologists. Check anesthesiology.org for an archive of this podcast and other related content.