Bias, Equity, and Reality: Issues When Using AI for ECG-based Diagnostics

Announcer: Welcome to Mayo Clinic's ECG Segment: Making Waves, Continuing Medical Education podcast. Join us every other week for a lively discussion on the latest and greatest in the field of Electrocardiography. We'll discuss some of the exciting and innovative work happening at Mayo Clinic and beyond with the most brilliant minds in the space, and provide valuable insights that can be directly applied to your practice.

Dr. Kashou: Welcome to Mayo Clinic's ECG Segment: Making Waves. We're so glad you could join us. Today we have an exciting episode planned for you as we discuss the issues around bias, equity and the reality when using these AI, artificial intelligence, ECG-based diagnostics. We're fortunate to have not one, but two expert discussions joining us today. Artificial intelligence has made its way into the world of electrocardiography. Just like any other medical diagnostic test, it is important to consider potential barriers and pitfalls when developing and applying these AI models. Today we'll be joined by Doctors Gari Clifford and Reza Sameni to discuss such considerations in the context of AI augmented ECG model development and how we can best deal with such barriers. Dr. Clifford is the chair of biomedical informatics at Emory University and a professor of biomedical engineering at the Georgia Institute of Technology. He trained in physics but then changed to electrical engineering and machine learning for his doctoral studies at the University of Oxford where he later became an associate professor. During his time at MIT, he focused on building huge public databases of medical data and he was recently elected as a fellow of the IEE for his contributions to machine learning applications and cardiovascular time series. He has been developing and applying machine learning models in the medical domain for over 25 years with the focus on open science through his leadership of the annual PhysioNet challenges, his application areas in cardiovascular disease, neuropsychiatric health, sleep and maternal fetal health particularly with marginalized communities with which he works with his medical anthropologist partner professor Rachel Hall Clifford and with whom he jointly runs and the Co-Design Lab for Health Equity. Now our other guest today, Dr. Sameni is an associate professor of the Department of Biomedical Informatics at Emory University. He completed his bachelor's degree in electronics engineering and he holds a master's degree in biomedical engineering and a double PhD degree in biomedical engineering and signal processing. Dr. Sameni was an associate professor and department chair of computer science and biomedical engineering at Shiraz University and a senior researcher at Grenoble Alpes University before joining the faculty at Emory University. His research interests include digital hardware design, statistical signal processing and machine learning with special interest in physics, informed modeling and analysis of biomedical systems. Dr. Sameni is a senior member of the IEE. He has contributed to training and research in the field of digital electrocardiography of adults in fetuses for more than 15 years. Doctors Clifford and Dr. Sameni, thank you so much for joining us today.

Dr. Clifford: Thank you very much for having us Anthony. It's a real privilege.

Dr. Kashou: Oh, well, it's an honor. Now, Dr. Clifford, I wanna start with you. And perhaps you can still help us understand, you know, when we are looking at these AI models and there's so much going on in this field and you know it better than I do. What are some of the key barriers that we should be, you know, informed about when we're building them from ECG data?

Dr. Clifford: Well, that's a long and complex answer. It's almost a PhD, more than the PhD itself but I think there are five key issues. First, there are large collections of data around the world that are just sitting in silos and not being shared. Most hospitals have millions of ECGs and they're being underutilized. And are often only representative of the patient population from where they were collected. So there are open source access databases and resources like PhysioNet pioneer the field leading us in posting free and open access to ECG data. But there's still very few complex large databases out there that really reflect a full spectrum of real world data. And I think this has really inhibited the innovation of ECG analysis. We're seeing some of that taking off these days but they're still on very monolithic data sets. But the time I think now is really ripe for disruption in this area. And in fact, we're actually assembling a new database. We've often posted lots of data but I still think you can level the same criticism at us. We've posted hundreds of thousands of ECGs and I don't think that's enough. We're currently working on posting a data set of around five to 10 million ECGs that we think will be much more representative of the population but it doesn't stop there. We need to go continent by continent and really assemble the true diversity of data. I think the second problem though even when we've solved that are other big issue is that the labels and the data are really noisy. And what I mean by that is that many ECGs are either only machine read so they're only as good as the algorithm interpreted them. And so if you train a model on that, it's just going to inherit the biases and inaccuracies of the original algorithm. Or even for expert labeled data, we only have a few experts who have labeled them. And for some classes of rhythms and diagnoses, the interrater and even intrarater, that's from one moment to the next for the same person. Those agreement levels can be as low as 60% which means that, you know, you can have 40% of your labels incorrect in a database. And that's a huge inhibitor for a field where you're expected to be 80, 90, 99% accurate. So having multiple algorithms and a multiple expert labeling data is a real issue. And how do we resource that? 'Cause it's a very computational and human resource intensive issue. So we're starting to do that through crowdsourcing combinations in competitions and also in online crowdsourcing platforms like our newly launched PhysioCrowd. And then there's another key issue with all of this which of course is bias which Dr. Sameni will talk about in a little bit. But if a patient population isn't well represented in the data, then you can't learn the issues that are unique to that population. So this could be unusual repolarization morphology, shorter QT intervals or all sorts of issues with different special populations from around the world. So having that diversity in your data is extremely important. The fourth barrier I think, training the algorithms is that the outcomes of importance are often not the metric that people use to assess the performance of the algorithms. So for example, you're really interested in treatment efficacy, long-term survival, mobility, mental health, quality of life, success of a surgery, these kind of things. And then not baked in when you are just assessing the error under the receiver operator curve or the accuracy or the sensitivity or the f1. These metrics are from information retrieval fields and they don't really exactly map to the things that we really care about in medicine. And then the fifth issue I think is the compartmentalization of data. So what I mean by that is that we have these ECG databases but the echocardiogram database for example is completely separate and it's very manual if you want to tie the two things together. Because of the way that we've specialized in medicine over the years, the databases have become specialized. And by that I mean isolated in their own silos with different standards. So sometimes it's almost impossible to match all your patients from one database to another even. So when we start to develop ways to combine these together in a more coherent way, that's true of every institution, then we're going to really be able to leverage

multimodal machine learning. So that's combining lots of different data modalities to make predictions which is, you know, why doctors are so good. Not just that their brains are very complex and have been trained well but they're also dealing with a lot more data than we feed to these machine learning algorithms. So that's my piece certainly.

Dr. Kashou: Thank you. And you passed. I think you've earned your doctorate just with that and what I heard were five different barriers. You know, the first being that, you know, we have all this data, right? But it's not centralized and they're in different silos around the world. We have label data as another problem, whether it's by the computer or what makes it into the clinical world. There's no kind of expert read, but as you rightly point out is that even expert reads have interrater variability. You know, is the presence of symptoms present when they interpret the ECG that it makes a difference. And so I completely agree. And, you know, how do we actually connect it all with the different hardware you mentioned? The bias that I'm gonna bug Dr. Sameni next about, and, you know, the barrier of outcomes which is, I'm glad you mentioned it because, you know, is it just the right interpretation or is the outcomes which how we perform a lot of studies, that's the important aspect. Those are the, you know, quality of life indicators, you know, mental health and those that you mentioned and then, you know, everything in a specialized silo. So I tried to summarize your thesis and hopefully I succeeded. But now Dr. Clifford talked about these potential biases and influencing factors when we consider developing these AI models, what can really be done to mitigate this bias in beyond the balancing of the data?

Dr. Sameni: So first of all, we can source large databases from around the world, from diverse populations with genetically different backgrounds. We should also do normalization for hardware differences. We know people around the world use totally different devices for acquiring the ECG. This can be done as part of the pre-processing stages of the data, which is extremely important. Things like basic properties of the input signal and system like the input bandwidth, cutoff frequencies, filters and things like that. They are extremely important and need to be standardized and compensated for during pre-processing. We can use the specific targeted enrichment for populations with poor access. For example, we can use paper digitization on the phone, something that we and others are working on to somehow leverage the hundreds of millions or more of diagnostic ECGs that are currently on paper and will be soon destroyed due to natural deterioration and lack of funding for preserving hard copy archives. We can also enrich data further for underrepresented classes using what we call physics-informed modeling. So cardiac anomalies are very diverse and have very unproportional prevalences. We can use synthetic data sets that look like the ECG and somewhat past it what we call the touring test. So if you show the synthetic ECG to a doctor, an expert, they would believe that it's more or less from a real subject. And these data need to be supported by the physics of cardiac wave propagation and can be used to synthesize arbitrary large datasets that can be used for training, data grading, machine learning and deep learning algorithms. At the classification step, we can use sort of voting between different models and penalizing for biases, something that we do every year in the PhysioNet challenges. And we also reported it in our latest ECG paper last year. So what we can do is we can vote between independently developed machine learning models and if we diversify the features and the machine learning algorithms, we would expect to get less biased results. But that's not enough. So in addition to diversifying the algorithms, we also need to modify our evaluation metrics. So generic machine learning metrics like PPV

accuracy and area under the curve and similar metrics, they are not enough. And we specifically need to design machine learning algorithms and to penalize for biases and building bias penalization terms into the machine learning algorithms that we use for ECG classification.

Dr. Kashou: And there's a lot there, you know, especially trying to deal with them. And I know you summarized it really nicely. I may have heard the physics-informed models. I don't know if that means making fake ECGs that are not human-based, but I think there's value there. You know, you rightly point out to these informed models that can help us in the penalty on some of these biases is important and perhaps a way to do it. So I'm excited to, you know, see more of that work and even look at your ECG paper. Now Dr. Clifford, can you expand, you know, a little more on what's meant by addressing biases much deeper than just balancing data?

Dr. Clifford: Sure Anthony. An example is think about the area of pain medication administration. So even if you develop a, let's say you develop an algorithm that assesses how much pain medication you should have for a certain level of pain you're experiencing. Let's pretend we have an objective measure of pain and we decide that the system also recommends a certain level of pain medication. The literature shows that there's an enormous bias from the prescriber point of view against particularly African Americans and particularly black women actually that they can tolerate pain more easily and therefore they get underprescribed pain medication. So I'm not saying that this exact analog in cardiology yet we have seen the 30th of 30 is shown that biases and differences do exist in different dataset in cardiology. And so, you know, there may be a bias in the way that we prescribe beta blockers, for example but we're not actually really very aware of this because we haven't done enough research into this area. And so I think there are all these hidden biases that we've not been measuring up until this point and we've suddenly become aware of these. So I think there's a huge amount of examination of the way that we practice medicine as a function of the algorithms that we've been using over the last few decades. Then at the other end of the spectrum is the equipment itself. You've got your clinician at one end and your patient interaction. And then right at the other end of the spectrum of where the data's coming from are the transducers that we put on the body. So we know that there's bias in the way that different pieces of equipment perform. During COVID there was a lot of oppressed about pulse oximeter being particularly bad about measuring oxygen saturation in very dark skin. This is Fitzpatrick four and five. And actually this work dates back to 2007 and even earlier where people have shown that many pulse-ox oxygens on the market don't detect desaturation in oxygen in the peripheral circulation. And it doesn't detect them as well in darker skin, I should say. And this means that, you know, you're less likely to get admitted for COVID, you're less likely to be given oxygen. You know, there's obvious really dangerous sequelae from that. And so there's an analog to this which is to give the pun. In the ECG we use different electrodes, we do different preparation of the skin. Some people just slap the electrodes on. Some people shave the skin before they put them on. You know, there's a lot of changes that change the background noise level. And these get baked into databases. So you could imagine that, you know, there's one database that has diagnostic bandwidth PCG and their power goes all the way down to North point North five hertz which means you can measure these changes in the st level in the ECG. And you have a high prevalence of ischemic or maybe you only take the ischemic patients out of that. And then you take another database of non diagnostic bandwidth where it's filtered from 0.5 hertz and down. And they don't have any ischemia in there. Your deep learning algorithm is just gonna learn that low frequency information means they're ischemic no matter

how much low frequency information is there. And so that's very, very important that we understand that the data itself, the way that the data's collected and which equipment we use will change the way that our algorithms are learning. And there's a danger that they can bake these errors into the algorithm itself.

Dr. Kashou: It's really fascinating. There's a lot in dealing with it and I'm learning a lot from just listening to you both. And I know you both deal with a lot of these large volume data sets and perhaps you could both address this before we end. You know, what advice would you give anyone wanting to apply AI on these large volume ECGs? Maybe Dr. Sameni, why don't you, do you mind?

Dr. Sameni: Sure. Take this one. So I can categorize it in three fields. One is data and analyzing the data. The second is the machine learning or signal processing design and the third is the interpretation. So for the data I would say select your data carefully, control for differences in sampling hardware populations and anything else that seems relevant to the target you're trying to classify your predict. For the signal processing and machine learning design, I would say looking to causal machine learning approaches, standard practices such as splitting the data into training, validation and test and keeping the test data totally unseen. So test data should only be tested once per algorithm and more than that results in sort of information leakage from the testing to that training phase of the algorithms. And at the interpretation level, I would say the the biggest danger in using modern AI in medicine is that it's so easy to fall into the trap of hypothesis free paradigm. So we need to try to be scientific and see if the predictions make sense. Perhaps use a resiliency map to see if the parts of the ECG that we are expecting really are triggering the classifier to light up. So for example, if the SC level is the most important trigger for a deep network in an atrial fibrillation detection algorithm, then something probably went wrong. Or if a bradycardia or tachycardia detector is not sensitive to the heart rate, like was there's something going wrong. So I would say the interpretation and really having some hypothesis and looking into the data and careful design of the machine learning pipeline are the most important factors.

Dr. Kashou: It sounds like those all three parts should be considered and I like, you know, avoiding the trap of the hypothesis free paradigm which we can sometimes certainly fall into. Dr. Clifford, any final advice for us with these sets?

Dr. Clifford: Yeah, I think it's important. Excuse me. Yes, I think it's important that we think about the end user. Who's gonna use it? It's unethical to design an algorithm and not specify the conditions under which it should be used. The FDA has this concept of labeling. So you say, okay, it can be used in the following situations but I think we need to be much more prescriptive than that because we need to be as open and transparent about where their algorithms were trained, what types of data they were trained on and where their limitations might lie. And so I'm not saying we should hide it or over-regulate it but actually you'd rather let people see under the heart so that they can test or tweak it. A nice paradigm is transfer learning, so you can put up these reference ECG algorithms with their weights in them like we do with standard image processing and have people build off the top of that knowing that that algorithm has a certain population baked into it in some way. And you can start from there and continue to train. And in

that way you are passing your diversity in your data on and allowing it to generalize from one algorithm to another.

Dr. Kashou: A few extra important points as Dr. Clifford mentions is, you know, making sure we know who the end user is, you know, making sure that we're considering them and also being transparent of what our models look like. AI augmented ECG models have shown tremendous potential. There's so much we are still learning and it's important for us to be cautious and transparent about the potential limitations of these models. We learned about some of the barriers involved in building AI models from ECG data as well as the potential biases that can exist in data sets and how to best mitigate them. These concepts are critical in the development of all artificial intelligence models if we intend to utilize them for patient care around the world. Doctors Clifford, Dr. Sameni, thank you both for joining us. You've really helped increase the awareness and help us better understand these important barriers when using AI to develop ECG-based diagnostic models. I know I've learned a lot and hopefully our audience is as well. On behalf of our team, thank you for taking the time to join us. It's been a true pleasure.

Dr. Sameni: Thank you for having us.

Dr. Clifford: Yeah, thanks very much. It's been a real pleasure. Thank you.

Announcer: Thank you for joining us today. We invite you to share your thoughts and suggestions about the podcast at cveducation.mayo.edu. Be sure to subscribe to a Mayo Clinic Cardiovascular CME podcast on your favorite platform. And tune in every other week to explore today's most pressing electrocardiography topics with your colleagues at Mayo Clinic.