Mind Your Data: The First Rule of Predictive Analytics in Clinical Research

In this episode of the IEEE Standard Association's "Re-Think Health" Podcast Series, Aaron Mann and Maria Polombini discuss how open data sharing is paving the way to access more quality, real-world and inclusive data to enable predictivity analytics to be more accurate, resourceful, and utilitarian in the world of clinical research.

Maria Palombini
Hello, everyone. Welcome to the IEEE SA Re-Think Health Podcast Series. I'm your host, Maria Palombini, Director of the IEEE SA Healthcare and Life Sciences Global Practice. This podcast takes industry stakeholders, technologists, researchers, clinicians, regulators, and more from around the globe to task. How can we rethink the approach to health care with the responsible use of new technologies and applications in such a way that can afford more security, protection, and sustainable, equitable access to quality care for all individuals.
You can check out our previous seasons of the podcast on ieeesa.io/healthpodcast. Here we are with season three: AI for Good Medicine, which brings a suite of multidisciplinary experts from around the globe to provide insights as to how do we envision artificial intelligence, machine learning, or any other deep learning technology, delivering good medicine for all?
We all want good medicine, but at what price, especially in terms of trust and validation in its use. As healthcare industry stakeholders, we're not looking for the next frontier of medicine. If it's not pragmatic, responsible and can be equitably valuable to all. In this season, we go directly to the technologists, the clinicians, the researchers, ethicists, regulators, and others about these deep learning technologies and what real and trusted impact can they have on improving outcomes for patients anywhere from drug development to healthcare delivery.
Will AI, machine learning, or deep learning cut through the health data swamp for better health outcomes? So a short disclaimer, before we begin, IEEE does not endorse or financially support any of the products or services mentioned and/or affiliated with our guest experts in this series.
It is now my pleasure To welcome Aaron Mann, Co-founder and Senior Vice President of Data Science of the CRDSA, Clinical Research data-sharing Alliance. Welcome, Aaron.

Aaron Mann
Hi, Maria. It's great to be here.

Maria Palombini
So today we're going to get to the basics- data. Why we need to make sure data going in will give us the benefit expected with predictive analytics and clinical research. Just full disclosure to our audience, the CRDSA is an IEEE ISTO Alliance. ISTO is the Industry Standards and Technology Organization. It is a global 501 C6 not-for-profit offering membership infrastructure and legal umbrella under which member alliances, such as CRDSA, and trade groups can stand themselves up as legal operating entities. For the world out there, you might not know IEEE does have this offering.
Let's get to humanizing the experience for our listeners. So, Aaron, tell us a little bit about you. You have a well-blended professional background having been a Program Leader at Genentech, a COVID-19 data-sharing Lead at TransCelerate Biopharma, and prior to that, a CEO of a big data analytics solutions company. As Co-founder of CRDSA, what drives your passion in working with data? What are you hoping to achieve with this Alliance perhaps you felt may have not been realized in prior roles?

Aaron Mann
I think the passion I have for data is what it is and what it represents. Fundamentally it's people, it's experiences.

Data scientists, sometimes we look at things as a series of data points, but for me, it the fact that those data points represent things that happen in the real world to people and especially in a clinical trial context, when you think a tough study say we're collecting about 2.6 million data points per phase three clinical trial, but each one of those is a unique part of a person's experience. I get passionate about what it represents. I think that drives a lot of why I get excited.
From a co-founding CRDSA aspect, really born out of frustration more than anything else. As an ecosystem we get really good in secondary use of data RIAs, data-sharing, talking about the challenges, the problems, what doesn't work and we're very eloquent in that. When we started talking, colleagues and I, people representing data-sharing platforms, academic research institutions, and sponsors, we shared this frustration of let's start talking about solutions. Let's get eloquent on solutions. Let's come together and form an organization that can solve problems that we can't solve in isolation as a single stakeholder or a single platform.

Maria Palombini
Absolutely. When I was talking to my colleagues in the world of blockchain we talked about blockchain for healthcare and blockchain for pharma, the real purists were like, we really have to talk about the data. He goes, we're not really getting to the core of this conversation. So, data seems to permeate all our technologies, no matter where we go.
The CRDSA is an alliance, obviously, and we all see these numerous amounts of consortium alliances that are being formed in many different areas of the healthcare domain. So how is CRDSA different? What is the vision of bringing this alliance together and what are the alliance's objectives?

Aaron Mann
We spent a lot of time talking to a lot of people before we decided to move forward, to make sure that we first understood the problem and how we might approach it, but also made sure that we were not duplicating anything that's already out there. It's important to understand that CRDSA is not a data-sharing platform. So, we're not a data repository, a data lake, but actually data-sharing platforms are members of CRDSA. So, our role is to represent the entire ecosystem. We have organizations like CPATH, Project Data Sphere , that are data-sharing platforms, data-sharing organizations, that are founding members, big biopharma companies, technology partners, CROs. We serve as the umbrella organization, looking for solutions that are common solutions to the challenges that we share.
If you really take a step back, the vision is how do we use the type of data that we have to dramatically improve the sharing and reuse of clinical research data and accelerate drug discovery. The easy way to say it, from an objective standpoint, is we want to make it easy to share and easy to use this data. Do we have enough volume going through systems and are we retaining high and updated utility and secondary views?

Maria Palombini
Absolutely. I think that's a well-blended mix of partners and participants you have in your group. So I think that gives it a really equal voice across the board.
Many times we've heard "what you put into it is what you get out of it." This might hold true for predictive analytics. I spent a good portion of my career, observing and researching the biopharmaceutical medical device industry and I never thought I would hear the words "open data-sharing" in clinical research or anywhere across the pharmaceutical value chain. We all have come to know pharma and clinical- heavy IP sensitive, regulatory complex, and the highest level of competition to get to the next blockbuster.
So can you share with us exactly or what is meant by open data-sharing in the world of clinical research and why this transformational shift over the course of the last few years?

Aaron Mann

We're in the middle of the transformation. I'm not sure we've actually shifted quite yet. We're definitely on that journey.
I think it is a mind shift that we've seen on the part of sponsors and research organizations. Data is not the new oil. That is something that you used to hear a lot more 10 or 15 years ago. It's not something that gets more valuable over time. The older, yes, the less valuable it is and it doesn't have any inherent value until you do something with it. I think one thing that's pushing transformation as senior leaders really getting that it's about how you use these data and that's where you're going to compete. That's your competitive advantage. Not the actual having of these data.
That leads to a second mind shift, particularly in clinical research, that this is patient donated data. It isn't something that's actually owned by sponsors, but sponsors are good stewards of that data. It's the patients that are coming into the clinical trial setting. They're donating their time and their data to further the science and reusing that it's an ethical imperative to honor the commitments that patients have made on the effort that they've put in to supporting clinical trials. That shift has happened.
I think the third, and in some ways, maybe most transformational, is advanced analytics - AI/ML. Because it requires big data, right? And as companies start building internal data marts, internal data-sharing capability, they quickly realize that, wow, no matter how big you are, you don't have enough data or the right data on your own. Even the biggest pharma companies, when you start looking at things like targeted populations in precision medicine will just, you need more. And so that recognition that you can't go it alone, no matter how big you are is something that I think we're just on the tip of the iceberg in terms of how deep that permeates organizations. But it's a shift that we've definitely seen accelerating.

Maria Palombini
Absolutely. I think you brought up a really valuable point because for so long we hear data is an asset, but for our accounting friendly people out there, data could be a depreciating asset.

Aaron Mann
I did a slide at a conference once in Las Vegas. I threw it up there like: data is not an asset. Data is an action. If I don't do something with it, it's just not worth it.

Maria Palombini
Exactly. It just sits there. Absolutely valuable insight from that point of view. I think this is a simple question, but I'm sure the answer is a lot more complex. Why has it taken so long for clinical researchers or sponsors of clinical trials to realize the potential of the reuse of the data from previous clinical studies? Perhaps the right question could be what exactly was prohibiting them from using it?

Aaron Mann
A little history helps context on this because the sharing of data in a clinical research, secondary reuse sharing is really a pretty new phenomenon. It started in 2013, at any scale, with a number of sponsors coming together with clinical study data requests (CSDR), but that was 2013. That was external data-sharing and data-sharing in that context, a little bit of an unfortunate term, right? Because it always looked to senior leaders, legal, your chief financial officer, like, well, this is us being altruistic and sharing out, but what do we get out of it?
In 2016/2017, you see the rise of that internal data-sharing efforts, and that really brought a sharp lens to what is it that we can do with these data? How should we be approaching it? I think that would have been accelerated, but there was a big monkey wrench that got thrown in about 2018 with the GDPR. It had this uncertainty around what data protection meant. And you saw a little bit of a slowdown where sponsors said, well, you know, I could share these data from previous studies, but am I taking risks when I do it? How do I understand that risk? How do I know what's acceptable risk?

And so that had thrown a little bit of a curve ball, but I think we now have the tools to really mine these data. I think the rise of AI, machine learning, predictive analytics, advanced analytics tribes has changed. Fundamentally has sponsors now thinking of themselves as data consumers, not just contributors.
But back to your question of what prevents sometimes the open sharing it's a chicken and egg problem. If their data scientists don't see enough volume to use an update of utility and external data that they can access and use, then it looks like a one-way street when it really isn't. It means that a company may not dedicate the resources that are needed to prepare to trials per sharing, make the policy decisions that are going to promote volume and utility.

Maria Palombini
Absolutely. We hear predictive analytics used across multiple industry domains. What kind of impact can it have on clinical research? Are we talking more efficacious clinical studies, more targeted patient recruitment, better meeting enrollment guidelines, all of the above or something different? Maybe you could share with us a case study where you have seen predictive analytics have a significant impact.

Aaron Mann
I think at some level it's all of the above, but the part that gets me the most excited is the creativity. What don't we know? When we combine different types of data, clinical research data with RWD, what new therapeutic pathways might be open or what new hypotheses do we generate that we can then go and test? So I think it's really exciting to think about the things that we don't know.
In terms of case studies that I've seen, there's a lot being done about earlier safety signal, identification, and classification. It's an important one. It's a place where early linkages can be subtle and therefore machine learning, for example, is particularly well-suited to making better predictive models based on early signals that may indicate later significant problems.
The other area that we've seen a lot of work being done, particularly around precision medicine is subgroups of population identification and the improvement of targeting inclusion, exclusion criteria, really trying to make the trials fit the use cases and being able to understand better how those responses will play out.
I think when you take a step back, most importantly, an outcome that we see is can we enroll fewer, but the right patients in trials? When we make trials more dynamic, terminate them earlier, where we save time and patient burden, when the predictive analytics are telling us that things may not be going the right way, conversely, moving them through regulatory pathway faster when we see there's good reason to hit that accelerator pedal.
I think all of these are use cases that have been done and are being done out there. You can't share the use case specifics problem in terms of being able to share broadly, but lot of work being done in the area and I think a lot of support within organizations for how this can play out and support their drug development process.

Maria Palombini
Those are really great outcomes. Everybody wants more inclusive and diverse populations, but targeted in their trials. So I think that could be a great contributor for sure.
We all know there's a difference between AI and predictive analytics. However, we know that they share a common challenge. It is this: if incorrect or dirty data goes into it than an invalid or erroneous outcome will come out of it. From your perspective, what's happening now with the data that is currently being used, that needs to be fixed and how can the work of the CDRSA eliminate or minimize these issues with the data before they're applied into these algorithms?

Aaron Mann
I think one of it comes back to a volume problem and an access problem. When I talk to AI advanced analytics companies, one of the biggest complaints that I hear is that we've built a really good tool, but all of us are

training our algorithms on the same sets that are publicly available, same datasets.
So I think there's a need for more diverse data and data sets that are ready for analysis and have high data utility. We use the word "data utility" at CRDSA intentionally, because it is clinical research data. The good thing is it is collected per protocol with defined outcomes, objective assessments, all of that. So it's quality data to start with, but it's going to undergo this transformation for secondary use. That transformation might be to protect patient privacy, it might be to protect IP, but it's going to go through something before it goes into, for example, an AI tool. That's the point where you can strip out utility. That's the volume bottleneck because it takes resources to do that.
So we're really working on both problems: volume and utility. The way we look at it is it's going to take movement on policy. Policy at data contributors, its sponsors, and from regulators to be able to bridge that gap of volume and utility and standards around what does good look like? I think all that we're doing is creating a better data model and data set utility, going into the powerful tools that are being created that power next-generation drug discovery.

Maria Palombini
I'm sure that would be very wanted by a lot of these tool-developers. I like to do this with all my guests and I call it the "think fast" question. When I mention "AI for Good Medicine," what's the first thing that comes to mind and why?

Aaron Mann
Creativity. What don't I know? What hypothesis did I not even think of testing, but because the system, the tool was able to interrogate datasets in a way that generated some new thoughts or insights, I'm able to develop a new way to look at a problem. That's the exciting part of this and the part that I get most excited about first-line.

Maria Palombini
That's opening pandora's box. Opening the unknown. What can we find out? For sure. We always hear a lot about ethics and AI. It's a big conversation globally. I think it's in every domain, not just healthcare. When we talk about ethics it's in the form of validated and responsible use in AI and machine learning for healthcare and I know that the CDRSA has some working groups on patient data governance, data protection, and data ethics. Why is this important in the scope of open data-sharing and what kind of baseline or blueprint are you guys trying to set for the industry to follow?

Aaron Mann
I think there's an essential tension in data-sharing. On the one hand, all these calls for open science. You hear those calls from WHO, NIH, share open line. But the same organization, like the UN can say, we want to open science and then say, privacy is a fundamental human right, which it is.
And so you have this essential tension between privacy, protecting patients, and open science. That creates this governance continuum in the middle. From a governance data protection standpoint, how you interpret as a sponsor, as a data contributor, how you interpret where you should be on that continuum, determines how much you're going to share, how much data utility it's going to have.
We have seen sponsors that have stripped out all adverse events and demographics from a contributed trial. Because they were being very conservative on the patient privacy side without balancing with the data utility side. That's the exception, not the rule, but it happens out there.
I think it also is around access. How easy is it to get to these data? Novartis is very public about their data42 project internal data mark, and they just published a paper and got to a point where their internal stakeholders can access their secondary use data in almost every case, it's an automated approval. In contrast, a sponsor I was talking to a couple of weeks ago where their researchers have to put in a formal research request backed by a

business case to access any part of their internal clinical trial data. So you have really different sides of that continuum.
For us, what we're trying to do or give people, sponsors, anchor points to say, this is a way that most people do it, it's not prescriptive saying you have to do it this way, but this is the balance of acceptable risk, acceptable IP protection that does the best job of fulfilling the ethical duty to protect patient privacy and the ethical duty to share openly and contribute to forwarding the science.
So we're trying to really create that blueprint or that anchor point that allows sponsors to have a comfort approach that they've got is one that is generally accepted best practice.

Maria Palombini
It's amazing that we still have this conversation about the tension between data-sharing and data privacy. When blockchain in pharma and blockchain in healthcare came out, they're like, this could be a potential viable mechanism for that and we're still here talking about it, but I think it's a very important, valuable point. In another podcast, they were doing a precision oncology study and it was the same thing. Trying to protect the privacy of the patients and what came out during the study was they actually had a suite of patients that they found other conditions in their data that they weren't even aware about. So they basically had to contact the doctor to tell them, listen, there's these suite of patients we use for the study that they have this condition and they may not be aware of it. Had the data being completely anonymized, we wouldn't even be able to go back to their governing physician and say that this problem existed. It's always that balance. Privacy is great and it's a human right, but I think you have to sort of balance the costs that potentially might come with it as well and I don't think anybody has that perfect answer.

Aaron Mann
I think you're right. The biggest frustration that I see technology companies in this space having, especially here in the US is that thinking GDPR and data protection at that level, it's a really sobering, eye-opener. You can't just reuse this clinical trial data as easily as you would think they should be able to. And so I think there needs to be understanding on both sides of what is acceptable data protection and sensitivity to that, as well as open science and bridging that gap. Again, it's a hard balance. There are a number of companies getting this right, or biopharma sponsors that get this really right. But it's still a big tension point.

Maria Palombini
Absolutely. We know there's a lot of vulnerabilities when it comes to patient data. We're talking about lack of security, every day there's either a ransomware attack or some sort of hack into a health institution. We have privacy issues, patient data governance structure issues. I know your group is currently working on the development of secondary use standards. So what sorts of issues are you guys trying to resolve through the development of those types of standards?

Aaron Mann
Our focus is on that transformation piece that we talked about. What happens during the transformation? What information is available about it? Right now, it's frustrating for data contributors because there's a lack of consistency across platforms, and often they are contributing the same trial across multiple platforms. It's frustrating for end-users/researchers because they don't have enough information about what transformations did or will take place to these data.
There's frustration around the sheer amount of data wrangling that needs to happen if you take trials from three, four, or five different sponsors and trying to pull them into one analytical dataset and find out that you have to do weeks of data management, just to harmonize it enough to start the analysis.
There's a real opportunity to have standards and accepted practices starting with just transparency. What are

you going to get when you get the trial? What is the supporting documentation you're going to get? What information you're going to get about what has been redacted down to the variable level. It really doesn't help when a sponsor says, well, we've contributed to trial, we had to redact some adverse events, but because of patient privacy, we can't tell you what they are. We've seen that. And that's just not helpful because now I'm not sure I don't know what I to know and it's really dangerous because I'm not sure whether that redacted adverse event matters to my research question could be central to it. And then if I'm using an irregulatory setting without that traceability and ability to know what happened from the original trial dataset to what a regulator is seeing, step-by-step you don't have visibility into that makes it very difficult to use it in a regulatory setting.
So our mission in secondary use standards is to start bridging that gap first by transparency on the transformations, and then moving through issues and challenges like data harmonization ultimately all the way through increasing the utility by having standards for how data should be transformed.

Maria Palombini
Fascinating. Wow, Aaron. You've given us so many great insights. I'm sure the shockwave was, data's not an asset. Let's call it an active ingredient for clinical research insights, but just for our audience, maybe you want to share a final thought, could be a call to action for data scientists or data ethicists, AI technologists working with the data who may be in this domain or interested in pursuing this area to support clinical research innovations. What would be your call to them or parting word of advice?

Aaron Mann
I think if you're with a biopharma company, if you're on the data management study side, be good stewards of the data that you have. Share it readily and well, and remember that you're competing on the analysis, not the data. If you're on the research side, the data science side, you're a biostatistician, understand that it's there, it's a competitive mandatory. Seek it out. Because from an organizational standpoint, there's no better reason that your organization, your company will share and participate than if you're biostatisticians your data, scientists are active users of these data. And I think on the other side, if you're an AI advanced analytics partner technology company, I think that to know is firstly, the data's out there, your specific client, a large organization may not know it's there, but it is. It is a real opportunity to push the competitive advantage of using particularly data external to an organization effectively.
So I think it's a real opportunity for the technology companies to be an agent of change and drive awareness and a mindset shift within particularly large biopharma organizations.

Maria Palombini
That's really important. Special thanks to you for joining me today and sharing these great insights.

Aaron Mann
Fantastic. Thank you so much. It's a great opportunity. Thank you again.

Maria Palombini
Absolutely. I could have talked to you on two other topics and take this podcast for a few more hours, but if you want to learn more about the CRDSA or how to become a member of the alliance, visit crdsalliance.org.
Many of our concepts in our conversation with Aaron are addressed in various activities throughout the Healthcare and Life Science Practice.
The mission of the practice is really engaging multi-disciplinary stakeholders and have them collaborate, build consensus, and develop potential solutions in an open standardized means to support innovation, ultimately helping to enable privacy, security, and equitable, sustainable access to quality care for all.
Activities we are in: wearables and medical IoT, transforming telehealth, decentralized clinical trials, mental therapeutics for healthcare, robotics for the aging. There are many different areas and they're all touching an

element of AI, machine learning, and the work they're doing. If you want to get involved, visit ieeesa.io/hls.
If you enjoy this podcast, we ask that you share it with your peers, your colleagues, or on your social media networks. This is the only way we can get these important discussions out into the domain by you helping us to get the word out. You can use #ieeehls or you could tag us on Twitter @ieeesa or on LinkedIn @IEEE Standards Association when sharing this podcast.
I want to thank you, the audience for listening in. Continue to stay well until next time.