What is multimodal data?

Genomics England transcript

**Naimah:** You may have heard us use the term multi-modal data, which is part of our Cancer 2.0 initiative. We have recorded this short podcast to help explain what this means. I'm Naimah Callachand and I'm going to be speaking to Dr Prabhu Arumugam, who is the Director of Clinical Data and Imaging at Genomics England. He's going to shed some light on this topic.

First of all, the million-dollar question, what is multi-modal data?

**Prabhu:** The programme is called multi-modal because in machine learning we have different modes. So, the analysis of an image would be one mode, or analysis of some clinical data could be another mode. The interesting bit here is what we want to do is take multiple modes, so DNA, X-ray images, radiology scans, and clinical data, and do all of that in one analysis or a single analysis. The reason we have come up with the term multi-modal data is that cancer is ultimately a very complex disease. Why it evolves, occurs and how it progresses are really complex and really difficult to understand. The general consensus now is that the more data you can bring into understanding the evolution of the progression of cancer, we think there is a real value to it.

**Naimah:** Can I just check what you mean by machine learning?

**Prabhu:** Oh, yes, that is an interesting thing. There is so much data in all of this kind of analysis, so if you think about it, DNA is thousands and thousands of basically letters. So, actually, one human being or even a series of human beings could not look at all of that data on their own. What you want is to write a series of codes to look for certain patterns or certain trends within your data. There is a whole series of terms that kind of go from artificial intelligence to machine learning, however you want to define it. It's ultimately not a human doing it, you are setting the precedent for a code, a system, or an algorithm to run through your data and find out new things or novel insights from your data set that probably we as humans wouldn't be able to, and that's a very superficial way of looking at it. Then there's much more complexity, so sometimes actually what you might say is, especially for images, you might ask a human to label specific parts of an image and say, this is a tumour, this is not a tumour, so there's even a degree of what you call supervision to the learning. What we want to head towards is what you call unsupervised. So actually, here is all of the data and what can this machine learning algorithm take away from itself and what can it teach us about that? So that is a really interesting bit for us.

**Naimah:** What are the issues with the current data that we have?

**Prabhu:** The first bit is building the library or collecting the data. We want to start with pathology, so most pathology labs in the country will take tissue from a cancer and place it on a glass slide, and what we want to do is digitise that. So, we have to actually scan it at a very high resolution on very specific scanners and that is quite a laborious task. It means actually collecting slides from NHS sites, sending them to a partner who is scanning them, digitising them for us, which is NPIC in Leeds, and then also having to collect radiology scans from NHS sites as well. So, the real complexity is actually bringing that data and bringing it in a format that is digitised and also independent of variability in different scanners and different sites, and so on. We want to really standardise all the images to start with. So, the first challenge is actually bringing that data in and then the second bit is actually starting to do the analysis. So, what are we going to look for? What is the kind of research questions we want to ask, and how do we make that data useful for us as well as? Which is kind of the second part of that challenge.

**Naimah:** And how many of these images are you trying to bring in?

**Prabhu:** At the moment, we are focusing on the cancer participants in the 100,000 Genomes Project, which is about 16,000 participants, and the aim is for each participant you are collecting glass slides. So, we estimate there will be about 250,000 glass slides or images, and an equal number of radiology scans as well. Numbers-wise, it is actually massive amounts of images, and they will also then have very structured clinical data and their genomes to go with that as well.

**Naimah:** What are the challenges of putting all this data together in one place?

**Prabhu:** There are lots of challenges actually, in terms of purely size. At a very simplistic level, each image is about one or two gigabytes. If you think about it, most kinds of iPhones have about 64 gigabytes. You could only really have about 30, or 40 images on one story to write and, actually, we are going to have 250,000 pathology images. So, the data storage itself is massive, and then when you add in the genome is about 200 gigabytes each. The vast storage of all of that is actually quite immense. We also want to do this in a very novel way, which is host it in the cloud. So, it is a very secure way of hosting data and storing data. Then the next thing is like running an analysis, so if you think about it, you've got hundreds of gigabytes of data for thousands and thousands of patients, it's vast, it's huge and if you want to kind of interrogate like that, it can take days and days of analysis to begin and run and interpret. So those are the challenges we have from what you might call an infrastructure and architecture perspective.

The other bit that we are also really keen on is ensuring patient confidentiality. So, all of our participants in our research environment have a Genomics England identifier. We replace anything that we believe is identifiable, NHS numbers, names, and date of birth, with solely this one identifier. That is how you track their genome, their clinical data and their images, and that is how you can link it all through, which is great for researchers and maintains that anonymisation of individuals. The challenges, obviously, is the more data you bring in, there is additional risk. You may pick up where scans were done, and so on. So, we are really focusing now on ensuring that anonymisation protocol continues, especially with more and more data that you can bring in and that's really important for us to ensure that for our participants.

**Naimah:** You've touched on a few things already, but what are the other benefits of bringing all this data altogether?

**Prabhu:** I think, first of all, it is the uniqueness of this as a data set. Around the world, lots of people have tried to pull together DNA sequencing and pathology images, and there is already a lot of understanding about the value of doing that. Actually, it is the sheer scale of what we can do. So comparative data set, something called TCJ and they have about 2,000 genomes and about 15 to 20,000 pathology images, so the sheer scale that we are going to go to is about 10 times that. I think the interesting bit is that we will also have a very structured clinical data set to go with that. So, it is a real understanding about when patients have had a relapse in their cancer recurrence, in their cancer, and the kind of chemotherapy and the treatments that they have had. So, that understanding that goes with the image and with the DNA sequences I think is really important there and it is, I think, the value that we can add to the research community to really understand the progression and evolution of cancer is unique.

I think our experience of what we have learned from DNA sequencing in particular is something that we can really add to that. Do not get me wrong, I do not think we as Genomics England are ever going to answer every single research question out there. I think the really important bit is that we can position this as a data set. Ultimately, it has been gifted by our participants for other users to answer lots and lots to research questions. I think that is a real push that we can advocate that this is a data set that can really expedite our understanding of cancer evolution and progression and actually how we can help participants down the line.

**Naimah:** What are the next steps? How will you look at the data once it has all been combined?

**Prabhu:** So, at the moment, we are trying to look in the next few years, so how would you use this data set from a research perspective, but actually where is the value for clinicians in say 5, 6, 7 years’ time? That drive is where we want to get to. I am not saying that we are going to solve that problem now. So, at the moment, a lot of it is testing what we are pulling from the public dataset. So, to TCJ, can we replicate that or expand analyses that have already been done elsewhere using our dataset? Then also we have got to build on that, so we have one of the world's largest whole genome sequencing of lung cancers, linking that to a radiology scanner, understanding the evolution tumours in upper lobes versus left and right lobes. It is really interesting how that evolves, looking at mutational changes, or you might call it like how cells look down a microscope on a digital image. There are lots and lots of things that we have got to start looking at, but the important bit for us is making sure that the data set is valid. We have done our kind of curation elements ourselves; it is standardised and it’s high quality, that is what we are expecting. Then really bringing users in to answer their research question, I think that is the next big step for us and making sure that data quality and so on, from our point, we have addressed it as early on.

**Naimah:** Do you think this approach will be used by others in the future?

**Prabhu:** I would like to think so. I think the important bit is that we have done something that is novel a lot of people are questioning about, is digitising pathology an interesting and useful avenue? And we have shown the value of actually doing that alongside whole genome sequencing. So, the drive for us is to show the value of what is possible and then actually the research from our data set will be unique. I think that will then expedite clinical sites to adopt digital pathology, adopt what you might call hosting data in the cloud rather than storing data in little hard drives in different places. So, there is a lot of value to what we are doing. I don't think I'm going to solve everyone's problems everywhere, and I really don't want us to say that, but we're just showing the art of what is possible, and I think we've learned a lot from that, and we'll be very happy to share what we have learned with others as well.

**Naimah:** Thanks, Prabs, for that really helpful explanation. If you have any other questions on multi-modal data, feel free to contact us at digitalimaging@genomicsengland.co.uk. Thank you for listening.