We are here today with Emma Thomson from the Centre for Virus Research in Glasgow. And she's going to tell us about the importance and challenges of integrating patient metadata with viral genomic data. Hello, Emma. Can you tell us a bit about your role?

Well, since the beginning of the COVID-19 outbreak, I've been very involved with researching clinical questions using data that comes from the genomic data that we generate through sequencing of the SARS-CoV-2 virus, and using that data to understand more about how well things like vaccination are working and how well treatments might work, for example, with different waves of the SARS-CoV-2 variants that we've seen from Alpha to Omicron.

So, can you tell us a bit more about exactly what is patient metadata or clinical metadata.

Well, we realised quite early on that it's very valuable to have additional clinical information, as well as genomic data from the virus that people are infected with. Metadata really refers to the data that surrounds the patients that we're looking at. And so that would include things like-- if we're looking at thousands of people, for example, who have been infected with sars-cov-2 in Scotland, it's quite useful to know whether or not those people also had risks for getting severe illness, how severe their illness might have been, what treatments they received, whether or not they were vaccinated.

These are all very key pieces of information that allow us to identify, for example, as the virus evolves over time, whether changes in the virus mutations are going to affect how well vaccines work, for example, treatments. And also understanding which groups of people are most at risk.

And from your experience working with SARS-CoV-2, which type of clinical metadata can be made publicly available but without compromising the patient privacy?

It's really critical that we can look at clinical data to understand these very fundamental questions about how well treatments and vaccines are working, for example, who is most at risk but without compromising someone's right to privacy and keeping within the legislation that surrounds that. And so we've been looking at new ways to integrate the clinical data with the genomic data.

And the way that we do that is to anonymize anyone. And in fact, what's really important using the genomic data, is that we can see differences between the different variants. So we know, for example, that Delta was more severe than Alpha. And we can see who in the population might be more or less at risk of each of these variants.

Unfortunately, as the virus changes, those sorts of patterns could change, and it's really critical that we keep watching that sort of information.

So, you did mention that there were systems in place to allow you to integrate these different types of data. Were these systems in place when the pandemic started, or have you learned anything new about how to integrate data now, as compared to what you were doing in the beginning? What were the challenges that you faced there?

So these systems for integrating COVID-19 or SARS-CoV-2 data with clinical data were not in place, and they had to be put in place. And they're actually put in place in different ways. In Scotland, we have a system, called The Safe Haven System, which allows national health service information to be matched to genomic data without compromising someone's identity or allowing us to analyse anonymized data sets.

Within the public health system as well, there are systems. And so we've been working with Public Health Scotland to integrate genomic data into patient records, actually, as well, so that that information is available. We will be able to use this type of data for other types of viruses. And in fact, also for things like multidrug resistance in bacterial infections as well.

And there's quite a wide scope now for upscaling the methods that we use to incorporate these more updated technologies that allow us not just to identify that the virus is present in someone, but also to identify that it's a virus that may have a mutation which could be associated with failure of one type of treatment or another and that we can tailor our treatments for people according to the sequence data.

Now when we first started doing research, we carried out a piece of research early on. And we actually had to get a network of treating physicians to go through patient notes manually. We now don't have to do that. We can actually use these anonymized systems to access that type of data.

And we can access it more readily and rapidly, as well, so that we're not looking back, but we're looking in real time at what's happening. And that's a very-- I think that is a very significant sort of step forward for us in terms of monitoring infectious diseases outcomes. And there is every reason why we could do that for other infections as well. And we should be doing it for all types of health care where genomic data is important.

It's important in cancer as well, for example, and other parts of medicine. And so, you can see that the scope is very wide. And I think one of the few benefits of having lived through the COVID-19 pandemic is that we have upscaled substantially the kind of technologies that we can use to provide people with better health care.

So, how is it that the genomic data can offer additional support when you get traditional epidemiological studies and vice versa?

Yes, so it's not always possible to collect epidemiological data from everyone. That's a very resource-intensive activity. And, of course, public health will do that, and with the various tracking programmes that have taken place in the UK, that has become much better.

But it was quite tricky at first. And I think the system was overwhelmed. And so, it wasn't possible to ask everybody who they'd been in contact with and which countries they've been in and so on. The genomic data can tell you a bit more about that. So you can actually see that the sequences, for example, were very similar to sequences which had been in Central Europe.

So we could infer that those sequences probably came in from Italy and Spain, in particular, and also other countries in central Europe. And that was really critical. I mean, as a physician at the time, we were treating patients and being told that we were only supposed to actually test people who were returning from China. And then, eventually, from the north of Italy, but we were always, at that point in time, a step behind.

And we weren't able to do the genomics in real-time, but we did it fairly fast after that month-- that first month of infection. Now we can do it in real-time. And so, we've really moved forward with using the genomic data. We started to see the advantages of using it.

So, following on from that initial exposure to the integration of these two types of data sets, what other questions have you tried to answer by integrating those types of data together?

So I think the most important other use of the data that we've been centrally involved in has been to identify vaccine effectiveness. So we can now, through the linkage of clinical metadata, we can see whether or not a sequence that's come from a virus from someone who's become infected is from someone who's had one, two, or three doses of vaccine or more. And we can also see what type of vaccine they had.

So we've been able to monitor how well those vaccines are working at not just preventing infection and test positivity, but also we're looking at whether or not people who've been infected with different variants are ending up in hospital or not and are dying of the infection. And so these really important endpoints we can now monitor at quite high scale and at a national level.

And these are incredibly important. I mean, it's obviously essential for us to be able to do that. And we'll continue doing that going on into the near future, I should think.

Is it possible for other researchers without a clinical background to actually have access to this type of data?

Yes it is. If researchers have a good question and they have appropriate stuff-- they're appropriately linked in with the health services, then yes, they can access that type of data. And in Scotland, it can be done through The Safe Haven System. I think that partnerships between public health agencies and the National Health Service and academic institutions is going to be really critical. And it has been very obvious in this pandemic how far you can move forwards when you have multidisciplinary teams and partnerships between those organisations.

Thank you so much, Emma. It was really exciting to hear about your work. Thank you for your time, sharing your know-how, your experience with us, and also for all of your work reaching the clinical and genomics interface. Thank you.