OC5_Jamrozik_data_sharing

[00:00:04.56] Hi, everyone. My name is Euzebiusz Jamrozik. I'm based at the Ethos Centre at the University of Oxford, and today I'm going to talk to you about the ethics of data sharing in the context of HIV phylogenetics.

[00:00:17.98] The basic idea here is that technical aspects of how data are collected, linked, analysed, and shared influence ethical considerations, such as our assessments of potential benefits, potential harms, how those benefits and harms are distributed, but also our assessments of how much these analyses respect people, respect their privacy, avoid stigmatising them, and so on.

[00:00:48.94] And this is part of a bigger discussion about open science. There's a fundamental ethical trade-off in open science, where on the one hand, open science principles are based on the idea that we want to maximise benefits.

[00:01:00.45] We want to have transparent data-sharing that improves the validity and how much we can rely on science and the results of analyses, and we can increase the possibility for different types of analyses and meta-analysis. And this, overall, can give more and more accurate results and ultimately improve the public health benefits of science.

[00:01:21.68] On the other hand, a lot of infectious disease research-- and today, I'll be talking about this in the context of HIV phylogenetics-- have the possibility of causing harm. People can potentially be identified by certain types of analyses or problems with data-sharing. That can result in stigmatisation.

[00:01:41.42] And there's a potential for misuse of data. So for example, in places where HIV transmission is criminalised, if police agencies get hold of HIV phylogenetic analyses that show who infected who, that might have criminal implications for some people.

[00:02:01.60] And a background consideration here is that there's many different types of data involved in phylogenetic analyses, and we can have different amounts or degrees of linkage of those data in order to do the analysis that we need to do.

[00:02:16.48] One thing we need to have for phylogenetic analysis is we need to be able to link individual HIV sequences with a database of sequences. In most cases, phylogenetic analysis is about the transmission of viruses between individuals, and so we need to be able to have a database of multiple individual samples to do that.

[00:02:37.70] It's often the case that we also link an individual's HIV genomic sequence with identifiers such as-- that identify who they are, where they live, when they were born, their sex, gender, sexual orientation, and so on. And that increases the identifiability of individuals, but also, it can improve the validity of the results of the analysis.

[00:03:02.19] When we link information about, for example, where people live to a database of sequences, that can show the spread of HIV, or indeed, other pathogens through populations, through communities and around the world that permits analyses like phylogeography analyses.

[00:03:19.95] When researchers and public health agencies have access to contact tracing data collected by public health agencies, that can also increase the accuracy of assessments of directionality of transmission, questions about who infected who, how high transmission clusters are spreading and so on.

[00:03:39.52] And these things can all be linked back to who individuals are to allow risk factor analysis, and also ultimately, say, linking people and their contacts to HIV diagnosis and care to try and produce public health benefits as a result of these analyses. So there's all kinds of different questions and options about how different types of data are linked.

[00:04:06.14] One key ethical principle here is so-called data minimalism, and you might also think it's also linkage minimalism. On the one hand, there's a scientific and ethical rationale to do the best possible analyses we can, a kind of data maximalism where we want to collect the most data, do the most complicated analyses that give us the best picture about what's going on.

[00:04:29.00] On the other hand, ethically speaking, we should only really link or have access to the data that are needed for that analysis and no more than that.

[00:04:38.16] So for example, data minimalism might reduce the risks of identifiability, stigmatisation, and so on by limiting the number of people's data that are shared or linked, by reducing the types of data that are linked only to those data that are necessary, by decreasing the level of detail about, say, for example, where exactly a person lives as opposed to a general fuzzy picture of their location such as a post code, for example.

[00:05:10.61] And also, we can also decrease the duration in time of linkage. We can produce a phylogeny-- a phylogenetic map of transmission in a population just for the purposes of the analyses, produce some results, and then delete the phylogeny after it's completed, and that prevents reanalysis and/or potential reidentification of people later on at least to some degree.

[00:05:37.96] But this raises questions about how we govern this type of research, or in some cases, public health activity. The data are held by different global research groups, different government and public health agencies, and in some cases, by different technology companies.

[00:05:53.08] And this leads to coordination problems about-- because different places have different regulations about how data are shared and handled, and we need appropriate gatekeepers or mechanisms of deciding which data gets shared, which data get linked, and for what purposes and so on.

[00:06:09.98] And there's also this tension between, on the one hand, research is increasingly international. It goes across, for example, high-income and low and middle-income countries. It goes across different states or different communities.

[00:06:22.18] And on the one hand, people want to have local ownership of data for obvious ethical reasons, but sometimes it might make more sense for the data to be held in other places. And so we have to kind of resolve that tension when we're designing systems to handle these kinds of data-sharing databases and analyses.

[00:06:43.13] So one upshot of this is that there's various technological ways we could share, store, and analyse phylogenetic data as a way of balancing the potential benefits and potential harms or risks. One option is so-called federated databases where the data are held by different institutions which are connected and there is an algorithm for moving data around for specific purposes under specific conditions.

[00:07:14.05] But one issue with this for phylogenetic data is that it's a very large amount of data in many cases, and there's technical barriers including the amount of computing power it would take to share the data that are needed under the appropriate security protocols and so on. And even if we had the computing power, it would also be very expensive.

[00:07:36.65] Another possible solution is so-called trusted research environments where data are deposited by different people into a semi-centralised system held by third parties who are often technology companies that run these environments. So we need to be absolutely sure that we can trust the people running the research environment and that we can trust the security on that environment.

[00:07:55.70] And the third option is the status quo of what people often do now, which is that researchers have a database, other people contact them to use a subset of the data, and there needs to be a clearly defined reason about-- reason and identification of which data are going to be shared.

[00:08:13.90] But this also involves significant costs. There aren't a lot of people who understand the database well enough to be able to do it, so there's staffing pressures. And it also raises some ethical tensions about, for example, how long we hold the data after a particular research project is completed for this kind of data-sharing which might maximise open science and so on.

[00:08:35.06] So in conclusion, different technical approaches influence the ethical features and outcomes of phylogenetic analysis. And one key principle here is a data minimalism, including linkage minimalism. And basically, this means that people should only have access to the data that are strictly needed to produce the results that have the optimal balance of potential benefits and potential harms.

[00:09:00.13] But it's a complex area and there's lots of unresolved issues that require further work, including potential new technical solutions. Thank you very much, and I hope you enjoyed the talk.