Transcript of the Webinar 4 questions and answers Topic: Data collection and management

Date of webinar: Wednesday 24th May

Introduction
Audio recordings were made of each of the webinars and transcripts were made of these. The following questions and answers are what was recorded in webinar 4 and are set out below. The only edits that have been made are to remove filler words (for example 'um') and repeat words. Some footnotes have been added which provide post webinar clarifications from the Spectrum 10K team.

Question: Schedule 1 Part 1 of the Data Protection Act 2018 talks of there having to be a 'substantial public interest' in respect of processing special category data (such as genetic data) for research. What public interest arguments does S10K have for collecting genetic data and meeting this legal requirement?

Answer (provided after the webinar): Special categories of personal data processed for research purposes are not relying on any of the 'substantial public interest' conditions – they proceed under the research condition for processing at Article 9(2)(j) of the UK GDPR ("processing is necessary for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes in accordance with Article 89(1) (as supplemented by section 19 of the 2018 Act) based on domestic law which shall be proportionate to the aim pursued, respect the essence of the right to data protection and provide for suitable and specific measures to safeguard the fundamental rights and the interests of the data subject"). See https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/the-research-provisions/principles-and-grounds-for-processing/#scd

So, in one sense the premise of the question is incorrect because it mistakenly suggests that a substantial public interest condition has to be met.

This said, there is a requirement to identify some sort of public interest (not *substantial* public interest) in the research for the condition cited above to apply. I would say there is a clear public interest in pursuing ethically reviewed scientific/medical endeavour to advance knowledge into and understanding of different conditions.

Question: The boycott Spectrum 10K campaign group, identified a section of the form were selecting no to having DNA used in future studies, or in databases means that you can't participate in the study, why is this sharing a requirement?

Answer: It's essentially because of how the study was designed. In general, when creating a study there are two sort of aspects that we need to sort of try and balance. One is the general idea in science that when funders try and fund studies, like this way, instead of collecting data, there is increasingly a need to sort of share those data's not in an ad hoc manner but make them available for approved research. And that's not you know, that's, that's there across multiple studies, for example, there's a study called the UK Biobank, which has collected DNA samples from 400,000 individuals in the UK and linked them to a lot of data resources, including questionnaires that people have completed. And this is all with consent. There are other smaller scale studies.

Usually, the way it works is that people apply, there's some sort of Data Access Committee, which looks at the applications. And then if those sort of applications meet a certain criteria, depending on the study, then the data is shared, fundamentally, the reason why this was required is because when we approached the Wellcome Trust to fund the study, the one premise of this was to create a resource called Spectrum 10K, where it's not just a research team, but other researchers can also use the data to address questions because the research team in Spectrum 10K might be small when they won't have the bandwidth or the ideas or the ability to answer questions, and they certainly won't necessarily have the expertise across all domains, so you might have experts in epilepsy who want to have, you know, access to data. So, I guess it's from that point of view, but when we designed the study, we wanted it to reflect the fact that this resource would be something where we would like the data to be shared with other sort of researchers. And hence, if people were not comfortable with that, and we fully appreciate that some people won't be comfortable with that, then perhaps this particular study, you know, was not something that would be sort of suited to their comfort levels or taste.

Just to add, we did make sharing with non-academic collaborators or private companies optional. So, because I think a lot more people would be less comfortable sharing with private companies, as opposed to just academic collaborators from outside of Spectrum 10K.

Question: Are the Spectrum 10K team aware of rulings by international courts which found that the indefinite retention of biometric data such as DNA profiles is unlawful. With Spectrum 10K saying now carry out long term processing of genetic data, with long term data not being defined. how can they guarantee that data processing is lawful in respect of such rulings?

Answer: Understand the concern when long term data retention isn't defined. Spectrum 10K is not doing indefinite retention of DNA data. So, after the analysis of Spectrum 10K key data, there would be an archiving of the data. So we are potentially looking at up to 25 years, but that is open to discussion. So, we would say there is not indefinite retention of the biometric data or of DNA data within Spectrum 10K.

On a practical note, it couldn't be indefinite because you do have to pay for storage. And you have to have a data manager, you also would have to have a data access committee for an indefinite period of time, and that would not be practical. So there's definitely an endpoint.

Question: And that might be something the co-production group think about?

Answer: Yes, exactly. I think what we're doing over here is not at some level, it's not unique, right. So, we are not building a model from scratch, what we have done is we have looked at existing models and tried to work with that. There are other long scale studies like I mentioned, UK Biobank, there's a study called [unclear in recording], all of which have got a clearly defined set of endpoints of various time periods and similarly Spectrum 10K also, as mentioned, would certainly not be sort of storing data indefinitely.

Question: So how will the data be protected in the years to come when the study is finished, and all the members are retired. So, what is the continuation plan for the data access committee and your database manager, and all those things?

Answer: We haven't got a plan for thirty years down the line. However, I would say that once the analysis of Spectrum 10K data is complete, the study will be archived, that means the data will be archived in a secure storage facility for 10 years, and then the data is deleted. So once the analysis is complete, in terms of how the project team will be managed over the long-term life of the project, that is something we would have to come back to you on. And let you know. And it's something we could discuss as part of the consultation.

It may be helpful to unpack the sort of timeline, right. Approximately as things currently stand, and there is scope for discussion about this, the data can be analysed for about 25 years, once the sort of study starts, essentially. And that would mean other researchers can also access the data and analyze it for approximately 25 years. And once that is done, and in this sort of 25 years, people are free to withdraw at any point. And their data can be deleted from future analysis. So obviously, where the data has been analyzed already, it's difficult to delete that, because some results may already be published. But from future analysis, that data can be deleted. And once it is done, the data is archived and during archive, no one can actually do any analysis, run any analysis, but it really is, for example, if there's some sort of audit in place and things like that, then you can sort of look into the archive for whatever is sort of needed. And after that it's deleted.

When it's archived, it's de-identified, so no, we don't keep personal information, because you don't need that. And under GDPR, you shouldn't keep that. So, at that point, when it's not been analysed, it would be completely de-identified and just for archive purposes.

In terms of data storage, personal information is stored separately from other data points that relate to a participant. So, when a person signs up to the project, there's a study ID that is assigned to their data, and their personal data is kept separately. So, it's pseudonymized.

Question: If the plan is to create a future resource or database of 10,000 autistic people's data to be used by other academics as they see fit subject to access committee approval, of course, then how does this not require indefinite retention of data?

Answer: Essentially, what we're talking about is what takes precedence is the sort of timeline that we sort of set over here. So, within that timeline, within 25 years, people have or other academics, you know, they can submit an application which can be looked at by the data access committee and you know, approved, or not approved, and then within that timeline they can analyse the data. Once the timeline is over, 25 years, that data will not be available for analysis to everyone. So, when you say long term, long term would also include 25 years that you know, under certain criteria that will be sort of long-term, but we're not sort of talking indefinite.

I also want to bring in the practical side of things. I know that there are other data, or at least genetic-based databases that are more indefinite, like the UK Biobank. And that's because they're funded to be set up for a longer period of time. But within the grant that we were awarded, we do also have constraints in how much can be done. And in terms of storing data indefinitely, that would be very costly, and it would not be possible within the grant. So, there are also practical considerations into storing data.

Question: When will the Data Access Committee be set up?

Answer: There was always a plan for there to be a data access committee that was built into the original ethics. The setup of a project is a very detailed process. And we will be speaking with the consultation about the structure of the data access committee, the policy and working process of the data access committee, the decision-making process, and we will be drafting a terms of reference for the data access committee that will be submitted to ethics as part of our document update to Spectrum 10K. I can't give an exact date of when the committee itself will be created, but it will certainly be built into the working group and to the ethics process. We expect that will be a discussion with a sponsor and ethics and the working group to confirm when the data access committee will need to be in place because obviously, we will not be sharing data immediately. That may be something that occurs after the project has restarted.

Question: The data access committee will include community members; will they be paid? How will data access committee members who are not there as part of another job be fairly compensated for their time?

Answer: I imagine they will be paid. We have to make decisions about this, and it will be part of discussions about how we run the data access committee.
Currently when we have people involved in any co-production for studies in the Autism Research Centre (ARC), at some level we do make payment and I think in some of our previous discussions, it was suggested that we would pay people for their time and expertise but obviously the details have not been finalised. And perhaps it's not for us to really finalise it, it needs to sort of happened through the co-production team and also depends on you know, the scale and things like that.

Question: Is there a process of vetting which future studies will get access to my genetic code?

Answer: The decision process and the discussion process about who can access particular studies will be filtered through the data access committee. Obviously, there will be a particular set of parameters that they must meet, they must demonstrate that they're in line with our values. And being totally clear and explicit about what those parameters are, it will be part of the working group to, to do that. When an academic collaborator applies to access the data, they would obviously only be able to access a subset of the data, they wouldn't be able to access all of it. And there would have to be a data transfer agreement in place. So, there are protections. We can't say at the current time, who's going to apply to use the data, because we don't know. But we can set the parameters around which data can be accessed or applied to be accessed. I hope that kind of helps make sense.

No one, no academic collaborators, no one will be able to access any identifiable data. They will only be able to access pseudonymized data, pseudonymized because they have been given, participants have been assigned IDs, essentially, and only the sort of data or the variables that people need to answer the question that they're interested in.

Question: Once someone has access to the data for an approved purpose, what's to stop them making a copy of it? Or using it or sharing it or stealing it? Once they've got access, how do you know that they're not going to misuse it?

Answer: The current system that we are moving towards, which was not really in place is not one where people can download. So, there are two different types of systems that are broadly used in research. One is where people can sort of download data onto the university approved servers. And, you know, generally, they can't download it onto their personal computers, none

of those things, because academic collaborators will need to download it onto a secure environment. And this is usually bounded by, you know, fairly strong data transfer agreements, which talks about, what are the consequences should they the data for purposes outside from what has been mentioned. I haven't really come across any sort of breach over here, because I think for researchers themselves, they are also bounded by the scientific reputation and things like that. And these breaches can be found out based on what is being published and things like that. That's one route.

The other route, which is getting increasingly popular is that people don't transfer data from one server to other servers, for example, to other institutions. But rather, collaborating, researchers can get access to the central database where they can sort of access the variables, only the variables that they want. So, within such an environment like this, it is close to impossible to download any data locally, you need to sort of receive permissions to download any data, what you can download are the summaries of results. So, suppose you run correlations or something like that. And you get a result of that correlation. You can download that. initially, when we started out, the second method where you can't download data was not available to us because it's quite difficult to build that infrastructure, because you want an infrastructure which can allow other researchers to access but an infrastructure where people can also run very computationally demanding analysis. But now I believe the university has developed such an infrastructure, which would mean that people would actually not be able to download data to their university servers, but rather they enter through a collaboration agreement to analyse the data on the university servers.

For example, if analysis is being done, one of the Principal Investigators [PIs] must have permission to access the data and work on it, they can't take it out, it's a very closed system, so that we have much greater control over the data.

Even through that system, people can't get access to it indefinitely. So, you will need to apply to access it for a set amount of period, say for a year or two year, however long people, researchers feel they need to sort of actually analyse the data, for example.

Another advantage of using a system like that where we control who has access to the data, but we also provide the environment in which people would analyse that data is that, if someone decides to withdraw from the study, we can withdraw them. From any further sharing of that data, we can withdraw them. We do have control over that side of things, which is something that is a concern that we've seen. And we've heard about. So we will have control over how well what data is being used when it's been used. And at at what point should that access be stopped.

Question: In whose interest may it be to stop the data sharing? Why is it seen as a bad thing to share findings of such big research that took lots of time and money? Why do you think that it is worthwhile? Why are you motivated to do something that's so difficult?

Answer: I'm going to answer this in terms of the value of data sharing. So obviously, science is a collaborative process, it's important that we're able to share findings to prevent replication of work again, and again and again. And also, it is a requirement by our funder that there is data sharing in place to ensure that the benefits of the research reach as many people as possible, and that it can be built upon. So, ultimately, with many of these things, it's a bit of a balance, right? I can sort of understand that people will have privacy concerns, there will be some people who don't want their data to be shared, who may want to come into a very specific study and they'll be completely okay to share their data for the very specific study, but may not be open to sharing the data for other studies.

There are different models in research, so there are research studies which seek to address a very specific question, where they sort of recruit very targeted individuals. Then there are at the other end, there are models where people sort of build you know, research biobanks or research datasets, which can enable other researchers to also address these questions. We have known for a fairly long time now that in research quite often, when you just analyse the very smallest sample sizes, you may get spurious findings simply because your sample is not large enough to identify some of the small things that you want to identify, or because the sample is not representative. So, there was a time in the 1990s, 1980s, when there were loads of these findings from fairly small samples, which were not replicated at all. And subsequently, what ended up happening was that a lot of money was funded towards developing these small samples, which people held very closely and did not want to share. And subsequently, no one could really replicate them. And the literature was filled with a lot of what we call false positives.

So the model that people are adopting now is to say, try and make it more open and transparent, so other research can replicate, or researchers can replicate your findings, and try and justify the sample size that you need for various findings that you want to investigate. And from our perspective, we know that the outcomes that autistic people want are quite different. Some people may want better support for epilepsy, some people may want better support for depression, so on and so forth. And these are very, very complex. So rather than perhaps designing just a small study, looking at epilepsy or something looking at depression, we thought it might be a useful move to design a broader study where we can collect information about multiple things. And we also do many of these things are interconnected. So, we also know for example, depression is linked to cardiovascular conditions, to also test the intersectionality between all of these different measures.

This is also linked to the previous question, in that we were aware to make this resource valuable so that other researchers could also access it within certain parameters, we would need to share the data with other researcher. One because it's good practice, so that other researchers can also access it. Two, other researchers may be able to replicate our findings. Three, also, because there may be other experts out there, there certainly are other experts out there who would have methods and ideas and techniques that we don't have, who can use this sort of resource. So, hence, that sort of balance, and hence why, if people are not comfortable sharing the data, in our original consent form we said, then perhaps you may not necessarily want to participate in the study because the study is built like that.

Question: How will you prevent this research from being used for other purposes in future which might result in the extermination or termination of autistic foetuses?

Answer: The honest answer is that this is a complex issue. And we can talk about what we are doing and then we can also talk about what the general field is doing and be a part of the general field, right. So, we have couple of steps that we are trying to enforce in that, we are trying to be quite stringent about who has got access to the data, which researchers have access to the data, whether they meet our values. I think a researcher has mentioned in one of the previous seminars that our values have been published at the Autism Research Centre [website] and clearly said that we are against selection of foetuses for these purposes. But then, there are certain things which are outside our direct control. And this is where the field needs to come together. When I say field, anyone working in autism genetics more generally, and psychiatric genetics more broadly, need to come together and have consensus statements. What is it they want people to do with the data and what is it that they don't want people to do with the data. And over here, encouragingly, one of the largest bodies, which is a psychiatric genomics consortium, alongside other bodies have come out and said, they don't support, selection of foetuses using genetics, for a number of ethical and technical reasons. And there are active efforts now to understand this a bit more and enter into dialogue so that there are greater regulations and restrictions. Ultimately, this is a broader societal issue, which we are very much in support of, to ensure that regulation and better restrictions are in place.

I also want to answer this from a technical point of view, in that, is it really feasible using our analysis where what we're doing is a genome wide association study? Is it really feasible to select for embryos? So, there is no way using genetics alone you could say someone is autistic simply because genetics is not everything, genetics explains only a small proportion of the likelihood for autism. So someone could have a reasonably high genetic likelihood for autism and they may not be autistic. And it's not just because of diagnosis, but they genuinely may not have autistic traits or features. Currently, as things stand, and in the future as well, the genetic scores that we have explain an extremely small fraction of the likelihood for being autistic, and we're talking about something like 1.2%. And it is very unlikely that this will

increase in any sort of large or meaningful way. So there are sort of technical limits to what can be done. And there are steps that we are taking as a research group, and also steps that the general field is taking to regulate this and prevent this.

Question: Not all medical records are electronic, so what about people, especially older people, who don't have electronic records?

Answer: Firstly, in terms of accessing electronic records, just to give a bit of background, how the process works is that we make an application to the national bodies in England, Wales, Scotland, and Ireland, to apply to access electronic health records. The dataset that they hold within each national body is slightly different, so you don't get the same dataset, you don't have access to the same record set in say, England as you do in Scotland, for example. So in cases where a particular parameter hasn't been electronically stored or retained, we wouldn't have access to the data. It wouldn't really be affected by age per se. Because we can go back to the beginning of the NHS, but it would depend on what records they have within the national body, for example, in NHS digital.

With electronic health records, there's an opt out for research provision, which people can say whether they would like their electronic health records to be used. Anyone through their GP, if they do not want their GP sharing their records with third parties of any kind, can opt out. And then it wouldn't matter if we applied to a national body, we wouldn't have access to data.

Question: With this study specifically, is it possible to participate without consenting to at least trying to access the medical records?

Answer: It wasn't the case initially, when we had built the study design, it was because co-occurring health was such an important part of the study, electronic health records would sort of provide an important source of information. But, we could discuss that in the cocreation group whether you want to sort of make that optional.

Question: So how will you ensure that all the processes involved are followed to keep data secure?

Answer: First of all, there are governance requirements in place, there is a structure in place, to ensure data is handled in a very particular way for research. And that goes across multiple levels. So, for example, we have sponsor oversight, we have ethical review, we have university policies and systems in place, we have internal policies for the ARC and through CPFT [Cambridgeshire and Peterborough NHS Foundation Trust] as well. So these provide layers of oversight and governance and data management on the project. Obviously, there are laws within the UK that you're required to meet and so, for example, we would never share

personally identifiable information with external researchers, for example, or anybody who's analysing the data. So there are procedures and structures in place within the governance and ethics structure of the UK that make sure that participants feel safe in studies.

Question: Do your research team have training in data management?

Answer: Yes, we have a data management policy, we have training through the university, we have training through CPFT and we also have a dedicated Data Manager.

Question: What's the withdrawal procedure going to be?

Answer: So at the moment, there are three levels of withdrawal, but we will be changing that. At the moment, anybody can withdraw at any time without giving a reason, that's absolutely fine. That's standard procedure for any study. If a person withdraws, they have an option in the current study, whether they want their data to be used in future research or not, there's three levels. But through the consultation, we want to discuss changing that down to one level, so the current more stringent level, which is the current level that anybody who has withdrawn in this project has been withdrawn at, which is that study data is only stored for archival purposes. So when a person withdraws, their sample is destroyed. Any analysis that is done up until the point of withdrawal is retained. So any data up to the point of withdrawal is retained. And that's a requirement under GDPR and governance for the project. But it will be stored only for archival purposes. What that means is that the data cannot be used in any future research, it cannot be shared, it's just there to confirm that the baseline statistics are correct, that we've met the parameters of how we've set up the project. It ensures that we can verify that X number of participants took part in the project, X number withdrew, these are so on and so forth, these are the baselines. For example, if a person decided tomorrow that they wanted to withdraw, their sample will be destroyed, they would be withdrawn, they will be taken off our database completely and there will be no data storage.

We can't use that data for any analysis, it won't come in the sort of database that researchers will have access to. But it will be transferred to an archive, which is essentially just a cold storage. And it's not something that a researcher would really have access to. But it's more if there's some sort of an audit or something like that, and to meet some of the governance issues. And even the archive, after 10 years all data gets deleted at the end of the study.