Centre for Personalised Medicine podcast

Series 2, Episode 6

Diversifying genomics

SPEAKERS

Rachel Horton, Gabrielle Samuel, Faranak Hardcastle

Rachel Horton
Welcome to the Centre for Personalised Medicine podcast, where we explore the promises and pitfalls of personalised medicine, and ask questions about the ethical and societal challenges it creates. I'm Rachel Horton, and I'm here with Gabby Samuel, and in today's episode, we're looking at diversifying genomics, a key aspect of ensuring that the benefits of personalised medicine can be accessed by everyone.

We're joined by Dr Faranak Hardcastle, Research Fellow at the Clinical Ethics, Law and Society group at Oxford. Faranak has just led a brilliant review aiming to identify key ethical, legal and social challenges in diversifying data. Faranak, please could you start by talking us through how you got interested in this area of diversifying genomics?

Faranak Hardcastle
Yes, thanks, Rachel. So, I'm a sociotechnical researcher, and I explore how technologies and societies shape each other and evolve together, and how we can intervene in this evolution to direct it towards a point where their benefits are equally distributed.

I was looking at the question of diversity from an AI angle, because there was a lot of discussion about how lack of data, or data that embeds inequalities, when they are fed into machine-learning algorithms, they might actually exacerbate existing issues? And there is a similar problem in genomics, which is a quite well-known problem that there is lack of diversity in genomic data. A lot of repositories and biobanks have data that are basically skewed towards individuals of European ancestry. And so a lot of other ancestral groups are underrepresented in these repositories. There have been on-going efforts to try to redress this problem, but there are ethical issues around these efforts that we should know before doing anything. The Clinical Ethics, Law, and Society research group that I'm part of is exploring similar issues as the field is shaping and so this was of interest to our research group and when we saw a call for a review on the ethical issues of diversifying genomic data, we got on to it.

Rachel Horton
Could you kind of explain to us why diversity is so important in genomic datasets?

Faranak Hardcastle
Sure. So we all have approximately 99.9% of our DNA sequence in common. And exploring that 0.1% that varies between us, can advance our understanding of how genetic factors

may contribute to disease or to protection from disease. And that's why a lot of times, scientists study DNA differences between individuals and groups. Those variants that are common in the population are usually unlikely to cause disease. And if they are rare, they may contribute to causing disease. But this may also very much depend on various other factors like social and environmental factors.

Another thing worth noting here is that there's probably more genetic variation within ancestral groups than between them. So for example, there are more DNA differences between individuals with North African and East African ancestry than between individuals with African and European ancestry. And so we, if we only study the data from individuals of European ancestry, then we may not get enough insights about the genetic variations in other ancestral groups.

Rachel Horton
So it's sort of a matter of having a good enough reference. And if you don't have a population well represented enough to know that something's common, you might then think it's rare and kind of make too many conclusions from that about whether it's causing disease?

Gabrielle Samuel
Can I ask a question about the examples because I know that you've always got quite a few really nice examples up your sleeve of where when you talk about these biases in either AI or genetics. Could you talk us through some of the examples where they could or have led to like, some health disparities?

Faranak Hardcastle
Yes, so I can tell you about a study by Harvard researchers, that was done on hypertrophic cardiomyopathy. The study initially used data that overrepresented European ancestry individuals but the researchers found that genetic variants were initially misclassified as disease-causing, whereas they were in fact common in individuals with African ancestry, and so they had to be reclassified as benign.

And as Rachel was saying earlier, it's sort of when particular groups are underrepresented in data, it's much more difficult to classify their variants as rare or common. And they end up either being misclassified or being labelled as variants of unknown significance. And so this, yeah, this example is just one of those examples that shows how easy it is to misclassify when we don't have enough data.

Rachel Horton
So the kind of consequence of it, like on the ground is people getting, like actively the wrong diagnosis or their family being tested for the wrong thing?

Faranak Hardcastle
Exactly, yeah.

Gabrielle Samuel
I mean, that's so interesting. And then, so, is most of this knowledge known? I mean, why…

you've conducted this review, what are you aiming to find in the review when you went to look for the types of ethical and social issues that you were after?

Faranak Hardcastle
Yeah, as you say, it's a very well-known problem, and it's been more than a decade now that scientists and clinicians have been calling for more diversity in genomic data. And there's been lots of efforts to try to redress this problem. It's just that the scale of the problem is so big that it's taking so long to progress with it. And so we were interested in understanding *why* this wasn't really happening. And that's how we got into understanding that actually, diversifying data is very challenging from a legal, social and ethical perspective itself as well. And so the review really wanted to understand what these challenges are. So yeah, so we just wanted to know, what are the ethical issues around the attempts to diversify genomic data?

Gabrielle Samuel
Well you've got to tell us more! What were the ethical issues that you came across in your review?

Faranak Hardcastle
OK, but before I get into the findings, I just want to say that the sort of search that we did for our literature review work has some limitations. One of the limitations was that most of the papers that we reviewed were from North America. And also another limitation was that our search mainly focused on underrepresentation that was based on gender, race and ethnicity. So that leaves out other underserved groups such as children, elderly, psychiatric patients, prisoners, and so on. And this is kind of like, this speaks to a problem about the attempts to diversify, which is that these categories a lot of times don't actually map to ancestral categories. So that's, that's one of the challenges.

So in terms of findings, we found that sometimes research practices can be exclusionary and this needs to change. One example, is approaches to recruitment or data collection that don't consider the cultural setting in which potential participants are situated. So, for example, for a group, group concern might be really important, but a lot of research practices may only focus on individual concerns. The literature suggested that practices need to have more cultural humility, which is often used to emphasize the importance of being reflexive and do active listening and taking responsibility for interactions on the side of researchers and research institutions.

Rachel Horton
So that sounds a huge issue to think about, could you just tell us a little more about what else came up?

Faranak Hardcastle
The second finding that I'd like to mention is the literature really emphasised the key role of coproduction in identifying and avoiding potential problems. So it's really important that potential participants are seen as active researchers and knowledge producers. And if we don't have such a mindset, then participant engagement, it's very easy for it to become tokenistic? And that can in turn risk exacerbating existing problems or creating new forms of

inequalities.

We also held a workshop as part of our literature review which helped us to complement the findings with some expert recommendations. And one of the things that came out of the wider literature review and expert recommendations was that there are lots of structural issues that we need to really keep in mind in efforts to diversify genomic data.

Rachel Horton
Please can you tell us more about those structural issues?

Faranak Hardcastle
One of them is that a lot of times, researchers might view data as neutral? But this ignores the fact that data and technologies cannot be separated from the social context in which they are created. And they tend to reflect our biases and social inequalities. If that's not kept in mind, then it's really easy to kind of make conclusions based on like shallow, sort of, you know, simplistic things that just come up in data.

The second structural issue was that these efforts need to really be contextualised, within the historical trajectory of structural racism and legacies of colonialism. And the third one was that classification and categorisation, as I was saying earlier, have political consequences, and they really need to be closely interrogated.

Rachel Horton
Could I just ask you a little bit more about you know, you were talking about data not being neutral. It'd be great to hear more about that, and what that means.

Faranak Hardcastle
Sure. So for example, during the pandemic, there was some research coming out and saying that there was some genetic susceptibility to COVID based on racial categories. And that was, to me, one of the examples of researchers going with the mindset that the data is neutral, but the reason, the *cause* of that, what was perceived as susceptibility to disease, genetic susceptibility, was perhaps more grounded in social inequalities. So that it's almost like, we need to be *more* scientific about these sorts of findings and interpretations.

Gabrielle Samuel
I think it's also about, right, what data we're collecting. So data isn't out there, we *choose* to collect, and what type of data are we choosing to collect? And why? And what does that say about our values? It's all kind of embedded in, I suppose, the data.

Faranak Hardcastle
Absolutely. And also the tools and methods that we use to measure things. They were all created by people at the end of the day, and those people, they came from their own perspective, their own experiences into that invention and application. And so it's all… it's all a matter of trying to contextualize all of these things that we use. And it's not about rejecting them and saying that, you know, they shouldn't be used, but it's about positioning them in the wider picture, to say that it's obviously comes from a particular angle and might not work well, when we're using it in a different context.

Gabrielle Samuel
It sounds like there are so many barriers and obstacles, I suppose for a researcher that wants to go in and try and do, I don't know, try to diversify their data in a way that is in line with ethically best practice. Did you come across… I mean, especially if it's at the structural level… did you come across any researchers that actually, I don't like to say this, but like, almost got it, right? Like, where you kind of read the papers and felt, yeah, that that worked well, or that… that had the effect it was supposed to have?

Faranak Hardcastle
Yes, we did find some really nice best practices that were from other countries that had been trying to coproduce genomic knowledge. But in the context of the UK, it may be that we need to really try and work out what works best for a sort of super-diverse society like the UK. So again, because best practices also talk about, you know, going to a specific community and just trying to get them engaged in research. But how is it to start from the beginning in a very diverse society? How can coproduction *naturally* occur? It's something that we haven't really, really explored yet.

Gabrielle Samuel
It sounds so complicated, right? Because when you talk about going out to communities and diverse societies, I suppose that leads us to the question of what… what is the community? And what kind of… even demographics are you looking for within a community? Because you said at the beginning, right, that the heterogeneity between communities is so broad – are you even looking for a community based on genetics, or socioeconomic, or, I suppose… yeah, it's a really interesting question.

Faranak Hardcastle
That's really a good question. I guess that's what I was trying to say, is that a lot of times the ethnic or racial categories that we have which are socially constructed don't actually map to ancestral groups. But what we know is that there's things like racism, or structural racism or structural inequalities, for decades have had biological effects on people. So, so yes, I mean, it's a really good question how you would first define diversity, then how you would define community? Some people, for example, define community based on geographical proximity. And some others talk about shared characteristics such as racial or ethnic categories, or shared lived experiences. But, yeah, it's a really good question. And it's something that it has to be determined in discussion with everybody and all those people that we're talking about. The answer is in coproduction, I guess.

Gabrielle Samuel
I remember in your report that you also spoke a little bit about diverse workforces, and the importance of going beyond diverse data. And it reminded me of something I read the other day about decolonising AI, and the needs… it's not just about the categories that needs to be thought about. But when we're thinking about the categories, it's *who's* actually conducting the research and what knowledge is being produced. And I was wondering if you could just talk a little bit more about that?

Faranak Hardcastle

Yeah, sure, so I think the categories have their own significance and importance in this discussion. But one of the things that we discussed in the report was that the push for diversity shouldn't be just about the data. It should be also about the sort of knowledge that is being made, and the sort of workforce that are in place, and also the disciplines that are getting engaged in the research. At the moment, there is a problem of lack of diversity in genomic workforce as well. And so, of course, these are all very much connected to each other, the more we have a diverse workforce, the more chances of having diverse data at the end, and so on.

Yeah, I mean, there's a lot of obstacles about how to cultivate a culture in the work environment, to sustain that diversity of workforce and discipline and so on. And so lack of diverse data is one of those things that you can't really tackle from one particular angle, and in silos, it has to really be thought through in terms of the bigger picture.

Rachel Horton
Thank you Faranak, it's so interesting… because, like, I guess the problem of underrepresentative datasets feels so kind of like, "Oh, we need to make that dataset more diverse", but I think this just beautifully illustrates how it's not as simple as just doing that, like, there's so many questions to consider, and so many things that get raised on this path to how we achieve genomics that is going to work better for everyone. If you were picking one message for people to take away from this podcast, what would it be?

Faranak Hardcastle
So it's really hard to convey all the complexity and challenges of this area in just one point. But clearly, there is a real problem that if we don't have representative datasets to inform genetic tests, it worsens the outcomes for people who aren't represented in those datasets. And this is an example of structural racism, having a system where the quality of testing you can access is so influenced by your ancestry.

But getting those datasets more representative needs to be a part of getting the whole enterprise of genomics more diverse, not the goal in itself. In fact, the key message from our review is that diverse datasets shouldn't be an endpoint in themselves. Just collecting genomic data from people with a range of ancestries doesn't address the diversity problem.

And this is because even if we have diverse data, that doesn't mean we have considered diversity in the true meaning of the term. To include diversity means thinking about diversity in terms of inclusion of underrepresented groups in all stages of the research process, ensuring that harms and benefits are equally distributed and to co-create knowledge so that the knowledge that is created is the knowledge that the diverse populations are interested in knowing, and ensuring that the benefits of that knowledge are fed back to a community or to that diverse population.

Rachel Horton
Where could we go to find out more about your work?

Faranak Hardcastle
So the draft of that review is now online on a preprint server. And we are at the moment in

the process of writing some academic papers from the review that hopefully will come out come out in the next year.

Rachel Horton
Ah brilliant, I'll be so excited to read those. It's such a fascinating field. Thank you so much for making the time to talk to us today about it.

Faranak Hardcastle
No, thank you for inviting me.

Rachel Horton
Thank you very much for listening to this episode of the Centre for Personalised Medicine podcast. If you'd like to find out more about personalised medicine and its promises and challenges, please visit the Centre for Personalised Medicine website at cpm.well.ox.ac.uk.