HOUSE OF COMMONS

# Science and Technology Committee

## Oral evidence: The right to privacy: digital data, HC 1000

Wednesday 23 March 2022

Ordered by the House of Commons to be published on 23 March 2022.

Watch the meeting

Members present: Greg Clark (Chair); Aaron Bell; Chris Clarkson; Dehenna Davison; Katherine Fletcher; Rebecca Long Bailey; Graham Stringer.

Questions 1-78

### Witnesses

I: Professor Andrew Morris, Director, Health Data Research UK (HDR UK), and Phil Booth, Co-ordinator, medConfidential.

II: Professor Christopher Holmes, Programme Director for Health and Medical Sciences, Alan Turing Institute, and Dr Melissa Lewis-Brown, Head of Research Data Strategy, Cancer Research UK.

III: Professor Aziz Sheikh OBE, Director Lead: Digitally Enabled Trials, Breathe Health Data Research Hub, and Director of the Usher Institute and Dean of Data, Edinburgh University, and Professor Chris Molloy, Chief Executive Officer, Medicines Discovery Catapult.

# Examination of witnesses

Witnesses: Professor Andrew Morris and Phil Booth.

Q1 **Chair:** Today we begin taking oral evidence in a new inquiry into the use of digital data, particularly in health research, and the questions of privacy that arise as a result of that.

To begin our inquiry, we are pleased to welcome Professor Andrew Morris to give evidence to the Committee. Professor Morris is director of Health Data Research UK, which is the UK's national institute for health data science. It brings together datasets from the NHS and other bodies for research into diseases and, we hope, their cures.

Joining us virtually, we have Phil Booth, who is the co-ordinator of medConfidential, which is a campaign for confidentiality and consent in health and social care data in particular.

Welcome, both of you, and thank you for helping us begin our inquiry. I will start with a question to Professor Morris. Let's start with the basics. When we talk about data, particularly health and medical data, what do you mean by data in your institute?

***Professor Morris:*** First of all, it is a pleasure to be here. I am Andrew Morris, a doctor and director of the national institute for health data science, Health Data Research UK, which is supported by the Medical Research Council, Wellcome and seven other funders. Our mission is to unite the UK's health data in all its multidimensional forms in a trustworthy way, to enable discoveries that improve people's lives, so it is about detail and scale.

When we think about health data it is not just about medical records and NHS data. If we are going to seize the opportunities of the fourth industrial revolution in a trustworthy way, it is about building the integrity of a trustworthy health data ecosystem that has the ability to link health data from medical records with omics data, which is biological data from genetics—proteomics—with pathology data, but also environmental and social administrative data, because it is often when one can link data in a trustworthy way across these domains that the most exciting discoveries and impact on human lives can be made. We prefer the term "health-relevant data", which spans multiple domains.

Q2 **Chair:** Thank you. Mr Booth, in terms of basics and definitions, you have heard what Professor Morris has said. Does that fairly reflect your understanding of the scope of the use of data and the questions of privacy that it gives rise to?

***Phil Booth:*** In large part, yes. There are obviously definitions of data in law, such as personal data and what is called special category data—about the physical, mental and emotional health of an individual—so it is important that we retain that understanding, because it is from those statutory safeguards that many things arise.

I very much agree that we should not just be talking about people's medical records. Genomic data, data derived from blood and tissue samples, and, increasingly, imaging data from things like X-rays and MRI scans that are being used by machine learning or AI, all raise distinct issues of their own that are not necessarily as obvious as looking at your GP or hospital record. As Professor Morris says, the linkage is crucial.

For example, UK Biobank has half a million volunteers who have consented to a range of uses, and yet 10 years after it was launched the Human Tissue Authority and the Health Research Authority had to go back and revisit consent because they hadn't looked and they weren't sure that there was necessary consent place for the sort of linkage they were doing. Things like the increasing use of whole genome sequencing start to create data that can clearly be to do with a person's health but can be used in many other ways as well. It is not just about linking the social administrative data to the health data. What we are creating in the course of healthcare, diagnosis and possibly even research can be used for other purposes, such as determining paternity and ancestry.

**Chair:** We will explore those purposes now. It is very helpful to understand that if we talk about health-relevant data, which is obviously quite an expansive definition, there is no fundamental disagreement about the scope. It is on the uses and consent that questions arise. I will ask my colleagues to explore some of the potential benefits of linking datasets in this way, starting with Graham Stringer.

Q3 **Graham Stringer:** Welcome, Professor Morris. Can you list the specific benefits and risks of sharing and linking health data?

***Professor Morris:*** The primary aim of linking and sharing health data has to be health improvement and improving people's lives. If you look at the best performing health systems in the world, they have two things: real-time information sharing across what I call journeys of care—primary, secondary and tertiary social care—which allows for optimal management; and the abstracted capability to research and analyse healthcare quality, and to innovate on those data at scale, which allows health systems to strip out waste, variation and harm.

In terms of benefits, health improvement is the main focus of activity. In my own world of diabetes, 20 years ago—locally, regionally and nationally—we in Scotland developed a real-time information system that linked primary, secondary and tertiary care to drive improvements. Within four years, we reduced amputation by 40% and laser treatment on the eyes for blindness by 46%. But we also, with the public, embarked on research activity—clinical trials and epidemiology, but also genetics. You have a virtuous learning health system where we use data for care, then abstract it for multiple benefits for research and innovation.

My final comment is that this is not new. The UK has been doing this for 30 or 40 years. It has come alive during covid, where we have seen a real convergence of care and research. When the smoking ban came in, did it work? Using linked health data, we showed that it reduced admission for

childhood asthma by 18% and admission for heart attacks by 19%, and 67% of that was in non-smokers. We have good evidence of using data. In clinical trials, we can also use data for long-term follow-up.

There was a study called WOSCOPS in the west of Scotland, involving 5,000 Scottish men. The initial trial took five years and cost £38 million. Using health data, they were able to look at what happened 20 years later, when they had all gone home, and the risk was sustained—there was an 18% reduction in mortality 20 years later. The benefits are pluripotential in terms of health improvement and clinical trials, but also in terms of epidemiology, discovery science and innovation at scale.

**Graham Stringer:** That is the plus side, but I also asked about the potential negative side—the risks of sharing and joining this data up.

***Professor Morris:*** I think that a key theme for us is demonstrating trustworthiness, because we—the research community—are privileged users of the public's data. The risks come down to demonstrating trustworthiness in terms of privacy and confidentiality. To be trustworthy, one needs to be honest, reliable and competent in how data are being used. I think we will come on to it later, but in terms of risk, we have to ensure that there are multiple safeguards in place to ensure that data are handled in a trustworthy way. There is something that we call "the five safes approach", including safe people, safe data, safe places and safe outputs, which supports the trustworthy use of data.

Q4 **Graham Stringer:** May I ask the same question about the benefits and risks of sharing your data with commercial entities?

***Professor Morris:*** Again, if we use covid as a reference point, from the diagnosis of a new disease in January 2020. The response, in terms of diagnostics, vaccines and antivirals, was a remarkable triumph of science. However, it was also a remarkable triumph of collaboration between Government, service providers—including the NHS—academia, and industry.

I would argue that, to address the pandemics of cancer, cardiovascular disease and mental health as we come through the pandemic, we need to define trustworthy ways of engaging with industry, so that we can innovate at scale. That is about being transparent, having trustworthy partnerships with industry, and being really clear around the principle of benefit sharing, so that we are transparent about who is accessing data, and how and for what purposes it is being accessed. Importantly, one of the principles and best practices of benefit sharing is equitability. To achieve that, we need to be much better at engaging the public, through constant dialogue, involvement and engagement, to work through what is regarded as being trustworthy in how we engage with industry.

**Graham Stringer:** May I ask Phil Booth the same pair of questions?

***Phil Booth:*** Because we advocate for patients—the people whose data it is—I think I would come at it from the basis that we do not know what is in any particular individual's medical history, and we certainly don't know

how they feel about that. However, we do know that that information is generally highly sensitive. There are things that people are reluctant to share with their doctor, or even with members of their own family.

I should say that I firmly endorse what Professor Morris said about benefits; I do not think that anyone is disputing those. However, the risks can be multiple and different for different individuals, should their information be used, misused, or abused. There could be embarrassment; as I said, people don't like to reveal certain things. There could be discrimination; there have been legal cases where someone's medical status has become available to people in the workplace, and that has led to discrimination against them. It could also, unfortunately, lead to things like scams, and unlawful targeting of the vulnerable. We have had to deal with a case like that with a company called Pharmacy2U.

That is why, as Professor Morris says, we advocate, and have done for many years, for the use of these trusted research environments, rather than the dissemination—the copying—of lists of individual-level linked data out to various entities. The main risk—I think we are in agreement here—is trust. We do not want to collapse trust, because without trust, we do not get all the positive benefits. That has been demonstrated by projects such as care.data in 2014 and one called GPDPR—the general practice data for planning and research—just last summer, and how the public responded to that.

This is not just about the theoretical; at an individual level, a loss of trust in the fact that what someone tells their doctor will be kept confidential may mean that people do not share information that is critical to their own health. As the pandemic has shown, that could also have been critical to the wider public health.

You asked whether there were particular risks with commercial entities. Of course, there is the same thing about trust and trustworthiness, but it is a pretty consistent public attitude that quite a number of people do not like the idea of profit-seeking organisations deriving commercial benefit from their data. In fact, it is one of the leading reasons why people might not be comfortable with sharing their information for medical research.

Unfortunately, historically the NHS and the pharmaceutical industry have obscured much of this sort of reuse through what are called information intermediaries—companies that service other companies, if you like, with data. Exploiting carefully framed terms like "cannot be exclusively commercial" still allows a commercial organisation to service the NHS and a whole bunch of other commercial companies, and, "for the promotion of health" goes along with a promise that patients' data will not be used for marketing or insurance purposes—but that still allows a lucrative market to thrive around what are called market insights and market access.

We believe in a TRE-only world, with proper governance and full transparency—commercial entities or anyone should be able to come and apply in the open for access to appropriate data for stated purposes, with intended value in mind and an equitable share in that value—but everyone

has to be following the same rules. We cannot have certain types of reuse lurking in the shadows. That is a necessary precondition to having the grown-up, properly informed public conversation that we want and that Professor Morris has referred to.

Q5 **Graham Stringer:** You mention insurance. Is there a real problem, or potential problem, of leak of this information to insurance companies, which would lead to individuals not being able to get life insurance or having to pay much higher premiums?

***Phil Booth:*** We do not have sufficient constraints on what insurance companies are able to do, necessarily. Over in the United States, a law specifically forbids the use of genomic information, for example, in certain contexts. At the moment, there is a moratorium, but it is not actually fixed in law that, if you have had a genomic test, your employer or insurer could not have access to that.

We caught an example of that happening in 2014, and I do not think that there has been much chance since of an insurance company getting your medical records from the central stores that we have—they would really have to hide their presence quite carefully to get access to that, through a process developed over time—but, again, we have found two British insurance companies using what are called enforced subject access requests. That is where somebody applying for insurance is invited to tick a box to allow the insurance company to make that subject access request to their doctor on their behalf.

That is routeing around the agreed report that a GP can give to an insurer. While I am pretty sure that insurers will not be able just to go digging in the information that is collected through NHS Digital, it is clear that they do want as much information as they can get, and sometimes they are willing to step up to and even over a line.

Q6 **Graham Stringer:** Can you tell us the name of those two companies?

***Phil Booth:*** I would not like to say now, because it is quite a few years ago, but I will certainly write to the Committee with their names.

Q7 **Graham Stringer:** Professor Morris, there are a number of companies out there that, if you give them a sample of your DNA, will tell you your family history. Other similar companies will claim to be able to tell you whether you are likely to get a particular disease, or that you are susceptible to diseases because of the structure of your DNA. When you talk about sharing different databases, do you think it would be useful to use the data of those commercial businesses?

**Chair:** Briefly, please, Professor Morris.

***Professor Morris:*** These are companies such as 23andMe, for example. You will be aware, and I will be very brief, that 23andMe have an existing relationship with GlaxoSmithKline to use that very large dataset to enable drug discovery, but that is based on a consented dataset where they have a relationship with the commercial provider. There is no doubt that the linkage of large genomic datasets will be transformational. Again, it is

about doing it in a transparent and trustworthy way. To agree with Mr Booth, the rules need to be the same for everyone.

**Chair:** Thank you. I am going to go to Chris Clarkson next and then Aaron Bell. I ask witnesses to be concise; your evidence has been fascinating so far, but we have a lot to get through. Please give concise answers to what I am sure will be concise questions.

Q8 **Chris Clarkson:** Professor Morris, I am going to start with you because I noticed on your CV that you were the dean of medicine at the University of Dundee. I am a graduate of that university, although unfortunately from the Scrymgeour building, not Ninewells.

I am going to ask a question about the law regarding data, specifically the uncertainties around data sharing. DCMS conducted a consultation last year where it made some suggestions about the future of data handling, including the elimination of article 22 concerning automated decision making, and introducing a list of recognised legitimate interests to allow for data processing. Do you think more legislation is required for data handling, or do you think it is an issue of training and guidance?

***Professor Morris:*** It is great to meet an alumnus. Briefly, a professor of law at the University of Edinburgh is a good friend of mine, and he says that legislation is not always the answer. I have two other points. I think it is about the caution, confusion and inconsistency of the current legal framework, which is actually very enabling for research. So it is around guidance and the application.

Q9 **Chris Clarkson:** It is about better using the tools we have.

***Professor Morris:*** Correct.

Q10 **Chris Clarkson:** Phil, can I take your opinion on that?

***Phil Booth:*** I basically share that opinion. To be equally concise, medConfidential thinks that lowering the bar for those who would misuse data, versus those who have to maintain high standards—like research—would tend to hinder the honest while empowering the dishonest. If we are talking about trust here, everyone has to be operating within the same rules and to the same high threshold. I am afraid that the DCMS consultation does not seem to appreciate or take account of the much broader and well-established framework of governance and ethics in both health and research.

Q11 **Aaron Bell:** Thank you, both. Following on from your and Chris's point that we should do more with what we have, we have a very fragmented data ecosystem and an awful lot of bodies are interested in this. How effective do you think the Government's arrangements are in ensuring that data sharing benefits society but protects individual privacy? If you have issues, what one suggestion for change would you make to simplify the data ecosystem? I will start with Professor Morris.

***Professor Morris:*** The ecosystem is very siloed and fragmented and there is a lack of consistency. For example, there was a paper in *The BMJ* last year about a group in London who wanted to look at the outcome of

people with congenital heart disease—people born with heart defects—and how their health was five, 10 and 15 years down the line. There were 47 sets of documents, 384 pages and the process took two and half years to link five datasets. That is an illustration of a system that is not efficient, is not productive and is not in the public interest. There is a need to pull together standardised approaches for governance, access, public engagement and transparency that people can subscribe to. Guidance from the ICO is beginning to emerge—the 14 March guidance on research provision—but we can do a lot more in enabling principles and best practice that can be adopted by the 219 trusts and 164 universities in the UK. The data ecosystem is vast. We are working in partnership to try to address that. HDR UK, under the leadership of Cassie Smith, is working with the ONS, NHSX and the devolved Administrations to shine a light on current practice, where the overlaps are and where we can agree on a standardised approach, just to try to reduce the complexity of the access and governance processes.

Q12 **Aaron Bell:** So in terms of a single suggestion, it would be about interoperability more than anything else, so that you can connect these datasets up better.

***Professor Morris:*** First, simplifying and streamlining the data governance and access processes in a very transparent way, with transparency around data use. Secondly, wider adoption of the five safes framework, and thirdly, meaningful public involvement and engagement every step of the way.

Q13 **Aaron Bell:** And in terms of picking one body to drive that, you would say that the ICO are the obvious choice.

***Professor Morris:*** The ICO can give data controllers confidence to allow data access and data linkage for public benefit. We are institutionally agnostic and a charity, and we are convening the major stakeholders across the UK to try to get some runs on the board in this area, because it is a sclerotic system.

Q14 **Aaron Bell:** Turning to Mr Booth, similar questions: how effective are the present governance arrangements, and what suggestions for change would you make?

***Phil Booth:*** By and large, I certainly take Professor Morris's point that it can be a very complex process for access, but in terms of the governance bodies themselves, it is not just the data protection regime we are talking about here. We are talking about a whole bunch of different data controllers that have essentially collected or created the data in their relationships with patients, and how that is managed. These are not just datasets floating around from nowhere; they are created in a context and under medical ethics, so there is a duty of confidentiality to all of this, not just a data protection duty.

This could be streamlined. What we have seen, unfortunately, has been so many different interests angling, lobbying and getting resourced for their particular set of uses that we have got this very disparate set of

arrangements—you know, a bit of empire building. If, when medConfidential was set up in 2013, we had been investing in a relatively small number of linked together or relating, mutually accrediting, trusted research environments—NHS Digital, one in Wales, one in Scotland, SHIP and SAIL, the one for genomics, the one for imaging—there would be a much simpler way for all this data to at least be managed. There would still be rules around access.

The most important thing that could change is that for every use of data, it is accepted that it has to be consensual, safe and transparent, and that we have consistent mechanisms for patients to be able to express their consent or dissent—which will affect whether their data is in a set or not—that these will be processed in five safes trusted research environments or equivalent, and that patients will be told how their data has been used. These are all entirely practical, certainly in a TRE world. As I am sure you can see, it would be far easier to have proper, transparent, trustworthy governance over a number of clear entities under the various relating regulators, laws and what have you, than to have what at the moment seems to be a sort of jostling between the silos as to who gets to use what.

Q15 **Aaron Bell:** Obviously, you want these clear principles to flow down. Who do you think needs to take responsibility for that, in terms of governance? Is that the ICO again?

***Phil Booth:*** At the moment, certainly around health, the Information Commissioner and the statutory National Data Guardian, which we advocated for, have to work in tandem, but the ICO is not competent—legally competent—and not responsible for the common law duty of confidentiality; the National Data Guardian is. In health research, there is the Health Research Authority and the Human Tissue Authority. I am not saying that there is one solution for all; I am saying that we can clarify the situation by being very straight with people and saying, "The data protection rules are coming from the ICO, and those are clear. Concerns about confidentiality are for the National Data Guardian. If you're using tissue or data derived from tissue, it's the HTA." These things are not difficult, but it can be very confusing because there are so many front doors.

Q16 **Aaron Bell:** You mentioned NHS Digital, which holds a lot of significant data repositories. The Government announced in November that it would be merged with NHS England, NHSX and Health Education England. Would you welcome that as a way of simplifying this, or would you be concerned?

***Phil Booth:*** No, and this is an important point, so thank you for asking that question. What the merger does is bring about an independent statutory body, which we think should be one of the key sources of data, but it breaks that statutory assurance by allowing NHS England to suck it in. NHS England is itself one of the large consumers of data, not so much for research as for what is called "planning purposes", which includes a host of things, and while many of them are perfectly legitimate, it is, to be

frank, a customer. You don't put the governance or the statutory safe haven into one of the customers.

There are clauses in the Health and Care Bill, which is just coming back from the Lords, that essentially allow NHS England by statutory instrument to take the powers that NHS Digital has. In the framing I have provided, it would be an extraordinary breach to allow an end user to direct itself to collect the data that it wants for its own—sometimes obscure—processes. I have to say that NHS England has not covered itself in glory with regard to transparency during the pandemic around its covid-19 data store and its attempt last summer to get people's GP histories.

Q17 **Aaron Bell:** Professor Morris, do you have similar concerns about this merger?

***Professor Morris:*** I focus on functions rather than structures.

**Graham Stringer:** Good answer.

***Professor Morris:*** There is no doubt that NHS England needs a very high quality and trustworthy data capability—

**Aaron Bell:** And governance structures to match.

***Professor Morris:***—and governance structures that are completely transparent.

**Chair:** Thank you very much. Before we go to Chris and Dehenna, I just want to say that we need to return to concision, I'm afraid.

Q18 **Chris Clarkson:** I will be as concise as possible.

**Chair:** I look forward to it.

**Chris Clarkson:** It's hard for a lawyer who's used to charging by the word.

Transparency in data regulation and particularly the GDPR relies on people understanding how their data is being processed. Phil, to what extent do you think current regulation of AI in data processing is fit for purpose? If you perceive gaps, what are they and how can they be remedied?

***Phil Booth:*** If we measure the culture and practice of CDEI and the opacity of DHSC's AI lab, it is not fit for purpose. We have a bunch of issues with AI, including really quite wicked problems around bias because the data we are looking at and is being used to train these models is full of historical discrimination. There is a clear need for not only data protection regulations, but looking at the accreditation and validation of these models themselves.

We—medConfidential—propose that we tackle this issue in health regulators by having what we call an analysis and input report. That is a certificate that shows exactly what data, what dataset, has trained what model, so that anyone, and certainly any public sector body, can know

when they purchase or commission AI or machine-learning services or products what the data supply chain has been. The equivalent of retailers' "This product has not been tested on animals" label, this certificate and this supply chain thing would help the regulators and show that no law had been broken and "no data subject has been harmed in the making of this AI."

We need to get more practical. There is a lot of talk about ethics and governance and a bunch of high-level stuff, but what is being done here is processing data. We can know where the data came from and whether or not it has been incorporated into a service or product, and as our ability increases to clinically validate some of these machine-learning things for live clinical use, that needs to be added to the certificate as well.

Q19 **Chris Clarkson:** Do you perceive any problems with the black box issue? You are talking about putting in ethical datasets and getting a certain set of results out, but we do not always know how the algorithm arrives at the conclusions it reaches.

***Phil Booth:*** Absolutely, and that is why I say it is one of the wicked problems. We keep up as much as we can with research in these areas, but I do not envisage a simple singular solution here, so we have said, as we often do, let's go with transparency all the way through the process, and if there is a black box at some point in the process, we focus all appropriate scrutiny and requirements on that and ensure the outputs are properly tested.

You may have a black box, but we are used to using relatively unknown things through the clinical trial process. One could still say, "This is a black box," like, "This is a pill," but there is a process you can go through so you can also still say, "But it's okay—this actually is safe." If we do not capture that but instead say people can start selling black boxes to public authorities, we have a real problem. We say let's get this certification bit going. We know we can look at the source of data, so if you are using a particularly badly biased dataset, obviously we should not be buying that for general use, but all these other things can be daisy-chained around. It is a practical and pragmatic approach to something that obviously will be used more in the future.

Q20 **Chris Clarkson:** Thank you, Phil. Professor Morris, do you have anything to add? I know you tend to deal with process rather than structure.

***Professor Morris:*** We need innovative regulation around AI. It is a challenge, because it is hypothesis generating rather than hypothesis testing. That needs very, very large, comprehensive, annotated and detailed datasets—detail at scale—and very large-scale compute. We need innovation in the regulation in a transparent way, so I subscribe to what Phil said.

Q21 **Dehenna Davison:** Professor Morris, what mechanisms are in place to protect privacy in data? Clearly, there is already some level of anonymisation often or pseudonymisation otherwise, but how do we strike the right balance between protecting people's privacy and making

sure that the data is still useful for research and has not been blurred out too much?

***Professor Morris:*** That is a good question. I will try to be brief. If data are anonymous, one loses the ability to link and the research innovation and public benefit utility, so there are trade-offs here. I would say: at every step of the way, being honest and communicating with the public, seeking their advice.

There is an emergent model of multiple concurrent safeguards. It emerged from the ONS work and the UK Statistics Authority—the so-called five safes framework. How do you authorise, authenticate and accredit people who are going to do research on the data? How can you de-identify and pseudonymise the data? How do you take the analysis to the data, rather than have a data dissemination model? As soon as you send data out the door, there is risk. How do you ensure that statistical disclosure control means that any outputs are anonymous, and how do you do this in a safe place that is completely auditable? What we are seeing is what I call power stations on a grid—these multi-tenant, secure and cloud-based trusted research environments, where projects and specific workspaces are auditable so that we know what is happening. Then we need to be honest with the public and say, "These are the benefits, these are the risks, and these are the safeguards that we are putting in place to try to minimise the risks as much as possible."

Q22 **Dehenna Davison:** And do you feel that the right balance is being struck there at the moment?

***Professor Morris:*** I think that there are examples of best practice in the UK and that the UK is being seen internationally as a place that is trying to strike a balance between trustworthiness and public engagement, and policies, technology and governance that demonstrate trustworthiness. But there is still more to do.

Q23 **Dehenna Davison:** Mr Booth, I know you have previously given some evidence on the point around anonymity and whether true anonymity can be achieved. Can you expand on that for us here in Committee today?

***Phil Booth:*** I agree with Professor Morris. For data that is useful for research—other than pure statistical research, because statistics are anonymous—the utility has gone. Let's be clear that we are talking about pseudonymised, or maybe de-identified, data. That is personal data in law. It may be inconvenient, but the simple answer is that, at individual level, richly detailed and linked data cannot be made anonymous. While I heartily endorse the safeguards that Professor Morris has just described—in fact, we have been advocating all of them—he missed out probably one of the most important ones: giving people a choice. Given that what you are going to be processing is personal data—we should just acknowledge this—we should do all the sensible and safe things to demonstrate that we are trustworthy enough to be handling people's data. None the less, we cannot assume that because we are doing that, we are good people and should be left to do whatever we want. We still need to give people a choice about what is being done with their medical records.

There are choices. There are what are called opt-outs as well as explicit consent—for example, for joining a clinical trial—but those opt-outs are still not being properly respected across the entire system. The national data opt-out is not being applied to all data releases. This point is critical. You can do all the things that we are talking about—trusted research environments, the five safes, taking an open safety model, and going to the data and doing the analysis there—but you cannot just do it because you think, "Hey, we've done this, so it's okay." You are processing people's personal data. While it is not always practical to ask for their prior informed consent, if people are given a general option to dissent and say, "I don't want my data used for these purposes," that must be respected by everyone, with some very clearly defined legal statutory exceptions.

Q24 **Dehenna Davison:** You have spoken about the importance of public trust in buying into user data. With regards to either opt-ins or opt-outs, do you think there is enough transparency and enough emphasis put on those? Often, we see them as a tick-box exercise that people tick through just to get to the next stage in a process. Do you think we need to go further there?

***Phil Booth:*** Yes, I think we absolutely do. We are talking about a health-related ecosystem. We have a national health service, which provides the healthcare of the vast majority of people in this country, and many opportunities to interact with people. What the Government have consistently refused to do is just write to every patient and be honest about the intention.

We've said this is the simple, obvious thing to do. They stuffed it up in 2014; they stuffed it up last year, and have done it other ways. Once we have in place all these trusted research environments that we have been promised, once the rules have been fixed, once the national data opt-out is made to work properly in all contexts, it would be a relatively straightforward matter to communicate to everyone once, to explain what the situation now is. They have heard enough bad stories—a drip, drip of bad reports in the press, with occasional outbreaks of attempts to get their data.

Let's fix the situation, get the dissent processes and other things in place. If anyone then does make any new initiative, it is on them to communicate that clearly, on the basic understanding that this is your right. You actually have a right to limit who sees your data, even in a direct-care context, which many people aren't aware of. That is around something called the summary care record, and should apply to what are called the shared care record, for direct care in this country.

We also have a right to opt out from secondary uses—what are termed research and planning. Unless we actually make this promise true, we are going to keep on losing trust.

**Chair:** Thank you, Dehenna and Mr Booth. Finally, Rebecca Long Bailey.

Q25 **Rebecca Long Bailey:** Remaining on the issue of ethics, how do ethical

considerations differ between the sharing and use of health data compared with other types of personal data? Starting with Professor Morris.

***Professor Morris:*** I will be very brief. Health data is particularly sensitive and is special category personal data.

***Phil Booth:*** Health data is created under medical ethics. It is 2,500 years since "First, do no harm." Professor Morris is correct that it is special category personal data. It is also confidential and exists within a professional, ethical environment, which includes academic research ethics, the environment for research. There are very different ethical concerns, but they are managed by some well-established ethical governance frameworks. Governments or new upstarts like the Data Trusts Initiative or whatever, coming in and saying they can fix things throws a whole bunch of history and good practice out of the window.

Q26 **Rebecca Long Bailey:** Thank you, Mr Booth, you mentioned that special category of confidential information that health data comes under. To what extent do you think the UK's health data ethics system, or lack thereof, is fit for purpose at the moment?

***Phil Booth:*** It could be much fitter. I go back to my point about consent. This is an important point, and although it may seem nit-picky and impractical, but we have to ask people, because their data is confidential and, if we don't respect that confidence, we have shattered trust. Although it might be easy for an adult now to do a national data opt-out online, it actually turns out to be not easy to do it for their kids. Families have to go through a paper-based process, which we found out last summer.

NHS dashboards are lovely and transparent, but they show that 50% of adults who opted out did not opt out their kids. There is a problem here. Again, very transparently, published each month, NHS Digital shows that two thirds of the data releases aren't respecting the national data opt-out. It is not fit for purpose and frankly undermines the UK's entire data ethics, if we do not and will not, once and for all, fix these very practical things around patient consent and dissent.

Q27 **Rebecca Long Bailey:** Professor Morris, to what extent do you think that the UK's health data ethics system is fit for purpose? What changes would you like to see?

***Professor Morris:*** The Caldicott principles are well established, including the eighth principle, transparency—so, I think increased transparency. Secondly, the opt-out process varies across the four nations, so we have inconsistency and that is an issue. Lastly, for some types of research, consent is intransitive. Onora O'Neill writes very well on this. It is very valuable, because it supports individual autonomy. In some uses of data, consent is a blunt instrument. Combined with authorisation, anonymisation and safeguards, it is a public dialogue around the benefits, then the safeguards and the risks, so I think that the role of consent in this is a key issue, as Mr Booth was saying.

Q28 **Rebecca Long Bailey:** Finally, on the opt-out system, Mr Booth said a few times this morning that it is not being enforced and that even when people are opting out, their information is being abused to some extent, in many cases. What should the opt-out system look like, if it is to be properly robust? Do the Government need to create more legal teeth, by way of financial penalties, for example, for those who are procuring or sharing data that has been opted out? Let me start with Mr Booth.

***Phil Booth:*** The implementation of the national data opt-out across all NHS bodies has been delayed by the pandemic. The date for people to expect it has been pushed back and back. That is, I guess, understandable. What it does not mean is that it can be put off indefinitely. What should happen is that it should be applied in a way that people understand it, which is, "I don't want my data used for planning research, or secondary uses", and that should happen, instead of behind the curtain a whole bunch of plausible legal exemptions that are found in order to continue to process the data. It is not trustworthy. It just does not come across as honest.

What we believe should happen is, because we have so much reorganisation going on potentially—moving from CCGs to ICSs, and we have all these other things going on out of the Health and Care Bill—a need for clarity. There should be, first, a single opt-out—I take Professor Morris's point about how it would be different in Scotland—that people may deploy if they do not wish their data shared even for their direct care. That is a very extreme case, but it is one that applies to some people and needs to be there for domestic abuse safety and all sorts of reasons.

There should also be a singular one that operates not as the national data opt-out operates right now, but more in the way that the current type 1 opt-out operates at your GP practice: if you say that you do not want your data to leave your GP practice for purposes other than your direct care, your data does not leave. They do not process it in some way and say, "It's anonymised, so it's okay", and they do not have a bunch of exemptions; they say, "It just does not leave." In that way, we can maintain trust. I think, over time, if that is in place, we would end up with a relatively small percentage of people using that opt-out. If we keep on coming back to say, "Okay, fine, we've fixed it now. We've done a bunch of other stuff, and you can trust us", we will just get another doubling of the people who opted out, as we saw last summer.

***Professor Morris:*** Currently, there are 3.5 million type 2 opt-outs. Opt-out is a very important principle, but it also has disbenefits, because we know that opt-out patterns are very lumpy—they are not evenly spread and vary twentyfold, and that could have public health implications.

Q29 **Chair:** Will you explain what you mean by "lumpy", please?

***Professor Morris:*** Sorry, it's not a very scientific word. Opt-out is not evenly spread across England. It is driven by age, gender, geography, GP practice—it's a 24-variation part—

Q30 **Chair:** And your concern is that that biases the sample.

***Professor Morris:*** It biases the sample. For example, if we had high opt-out in—when we're looking at public health issues—Camelford, Salisbury, Sellafield and there was differential opt-out, we wouldn't be able to understand the underlying aetiology of those events.

**Chair:** Thank you very much indeed, Rebecca. I thank both our witnesses, Professor Morris and Mr Booth, for a fascinating series of answers to our initial questions. We are very grateful.

## Examination of witnesses

Witnesses: Professor Christopher Holmes and Dr Melissa Lewis-Brown.

Q31 **Chair:** I am now going to turn to our next panel of witnesses, both of whom are appearing virtually—I can see them in front of me. They are Professor Christopher Holmes, the programme director for health and medical sciences at the Alan Turing Institute, which is the UK's national institute for data science and artificial intelligence; and Dr Melissa Lewis-Brown, the head of research data strategy at Cancer Research UK. Thank you both very much indeed for joining us. Perhaps I can start with Dr Lewis-Brown and ask you to set out briefly what the advantages are of sharing datasets, particularly as you come from an organisation charged with finding cures for various types of cancer.

***Dr Lewis-Brown:*** First, I will just give the context that I am speaking from a number of perspectives in my capacity as the Cancer Research UK representative. We are a major funder of health research. We fund hundreds of millions of pounds-worth of cancer research each year. This includes data-enabled research such as with the data that we are talking about today. We are also patient advocates: we have an 1,800-strong network of people affected by cancer and five patient panels, with whom we deep-dive into issues such as those that we are discussing today. Also, we are a user of health data: we have a cancer intelligence team that analyses the sorts of data that we are talking about today, to track cancer outcomes in the UK and work towards service improvement. Data is extremely important for all of that.

Some of the benefits of data sharing for cancer research are that we can better understand individual cancer risk and can speed up diagnoses; and the kind of research that we can do using this sort of health-relevant data can drive quality improvement and optimise services. Just to take one of those as a specific example, early diagnosis is so important in cancer—it's critical. To give an example, the five-year survival rate for patients with colorectal cancer is 90% if they are diagnosed at stage 1 and only 10% if they are diagnosed at stage 4. So a major example of the benefits of health data sharing in the cancer realm is that it allows us to find new ways to detect cancer earlier and make diagnoses earlier, which leads to life-saving interventions.

Q32 **Chair:** Thank you very much indeed. Professor Holmes, in your case, I am

thinking particularly of the application of artificial intelligence in the analysis of these datasets, which is something that the Turing Institute has particular expertise in.

***Professor Holmes:*** Yes, thank you. Perhaps I can give an illustration or anecdote. Imagine a GP practice of, say, 10,000 patients. I walk in with a particular condition, and the GP would like to look at what worked well on patients like me—did drug A or drug B lead to better outcomes? They might search their GP records and, because there are 10,000 people, they might find someone who is a little bit older and maybe a little bit lighter, but only a small number of people would be close to matching. If we extended that to a GP practice of 60 million people and started to look for patients like me, we would get very high resolution matches. In one sense, what AI is about and scientific research is about is matching; it's about matching individuals. So, if we can combine at the scale of a national level, we can find highly granular matches. This doesn't just look at my age, my sex and my weight; it can also match on genomics and individual information coming, say, from a smartwatch. It is the granularity that gives much greater precision to provide clinical decision support which, combined with the expertise of the GP, can lead to better treatment recommendations.

The other very important aspect is on the research side, which is that combining data at scale allows for discoveries—to find what are things underpinning, say, genetic risk for chronic diseases. It is an essential component. It is a number game. The bigger the numbers, the better the recommendations we could make and the more robust and reproducible scientific findings we can support.

**Chair:** Thank you. That is extremely clear.

Q33 **Chris Clarkson:** I want to talk to you about some of the issues with lack of legal certainty in data sharing. Dr Lewis-Brown, to what extent do you think the Government's data-sharing proposals address the legal complexity of data sharing? I would like to drill down into what potential dangers, if any, the Government's approach to streamlining data sharing legislation will have. As a subsidiary to that, how important is the EU's adequacy agreement on data for Cancer Research UK?

***Dr Lewis-Brown:*** We welcome the Government's appreciation of the scale of the challenge that these legal complexities present to data sharing and health research and the extent to which they have looked at legislative reform as a solution. However, we think that the proposed reforms will not have the desired impact. We do not think they will address the legal complexities.

Unfortunately, the reality is that the legal framework around data sharing is complex, and there is perhaps not that much that we can do to simplify that. These proposals certainly do not streamline those arrangements. We suspect that they may actually lead to further confusion by implementing these changes.

We think that a more effective solution is to provide better and stronger training, guidance and support to those whose role it is to interpret those complex legislative arrangements when it comes to data sharing. We have seen a lot of inconsistency in the interpretation of the existing data sharing legislation, which in some cases has led to unclear and inconsistent decisions, and that is often steeped in a disproportionate risk aversion. We think that, at best, they do not simplify the legislation, and that a better solution is stronger training, support and guidance. At worst, we feel that they undermine the EU data adequacy agreement.

To your follow-on question specifically, the agreement is incredibly important to Cancer Research. Members of the Committee will know that data adequacy is a term that the EU uses to describe countries that provide essentially an equivalent level of data protection to that in the EU. It is on that basis that Cancer Research UK funds a very large number of clinical trials to develop new treatments for cancer types. A third of the trials we support involve an EU member state. If that adequacy agreement was undermined and we could not rely on it to carry those trials, we would have to stop them—50 or 60 trials at the current time.

One of the reasons that we enter into international collaboration for clinical trials in cancer is to find new treatments for rare cancers, because as a small country we do not necessarily have a sufficient number of people with any particular rare cancer to be able to do research on our own. We need international collaboration so that we have a large enough pool of patients to be able to be able to do statistically meaningful clinical trials.

Actually, an example of a particular patient group who are afflicted by rare cancers is children, so there is chance that this might disproportionately affect the development of new treatments for children with cancer—those who are the most vulnerable in our patient populations.

Q34 **Chris Clarkson:** Thank you. That echoes the comments of Professor Morris on the last panel. Would it be fair to say, especially when you are working internationally, that changing these guidelines would have a deleterious effect on the quality of the data and the ability to match it to exterior sources?

***Dr Lewis-Brown:*** Absolutely.

Q35 **Chris Clarkson:** Can I quickly follow up on legal compliance costs for data sharing and how they impact your research and innovation?

***Dr Lewis-Brown:*** There is an economic cost to legal compliance. The fundraised money that we have to spend on legal compliance could otherwise be funding cancer research directly. That is certainly a consideration. With reference to the last question, if we fall short of the data adequacy agreement, we will need to spend a considerable amount of time and resource on trying to develop alternative legal mechanisms, so that we can enable those clinical trials.

The main cost is to cancer patients. Legal compliance takes time, and the resource that's pumped into legal compliance reduces or slows life-saving

cancer research. The one thing that cancer patients do not have in abundance is time.

Q36 **Chris Clarkson:** Professor Holmes, can I turn to the use of AI in data, please? In the previous panel, I asked Phil Booth about how he felt the current regulations on transparency of handling of AI data worked, and particular issues around black box problems. I want your thoughts on how the current regulation operates, and whether it is fit for purpose and takes account of the role that AI could play in data research. Can consent be built into the use of AI when it is processing an individual's data?

***Professor Holmes:*** On the first question about regulation, it is certainly the case that AI introduces new complexities and challenges for regulation, in particular the more opaque versions of AI. We should be aware that this is a continuum, so there is not a discrete classification of what is an AI system and what isn't.

The labelling of a system as being AI or not is problematic. What we are talking about is algorithms that can learn from data, as Professor Morris called it, hypothesis free. We present it with data and ask the algorithm to perform a search in order to help us make, say, a clinical decision.

I would say we need agility in the regulation as we move forward. Because we are in a time in AI of incredible discovery and innovation. There is no doubt that the AI algorithms and systems of tomorrow will look different from what we have today.

As we solve the data linkage problem and start to bring increasingly large datasets together, new algorithms will emerge that will have potential for strong public benefit. On the question of regulation, I would say that we need agility within regulation but, of course, with all the safeguards around transparency and trustworthy that go alongside that.

Q37 **Chris Clarkson:** To follow up on that, if we are asking algorithms to make decisions, would you say it is important that we have at least a reasonable understanding of how they work, when we get to those black boxes that Phil Booth mentioned?

***Professor Holmes:*** People would typically call it human in the loop AI, to assist human decision makers, to free them up from simple decisions and allow them to concentrate—for experts, say clinicians, to be able to concentrate on those cases where they have best use of their skills. It is certainly the case that you can adapt AI in a sense to be explainable or not. Putting down constraints on explainability is certainly one way of counteracting that.

Q38 **Chris Clarkson:** You see its role more as augmenting the existing process, as opposed to replacing it.

***Professor Holmes:*** Absolutely, with greater agility. Agility needs to be woven into the regulatory system.

Q39 **Katherine Fletcher:** Thank you for your time today, which is enormously appreciated. I have to cough: Professor Robert Sapolsky once told me,

"Congratulations on being a recovering biologist." I thought that was a good way to put it in this job.

I have a deep, underlying concern that, with advances in technology, it is not possible to anonymise DNA. We've got cases in America where people who have perpetrated crimes are being caught because their cousin wanted to know their antecedents and where they came from, via an online commercially available test.

I am interested in your perspectives on this idea: if I have a heart attack tomorrow, I am happy to share that data and all the surrounding information, to help future research. But Professor Holmes, in your opening remarks, you mentioned the idea of genomics anonymised being an enormous opportunity area. If I give DNA and permission, I am in effect giving permission for my family, nieces, nephews, cousins. I wondered how you square what is clearly quite a difficult circle. Because there are a lot of benefits to be had. I will start with Professor Holmes and then come to Dr Lewis-Brown.

***Professor Holmes:*** That is a fascinating question. I think genetics play a particular role in revealing recent ancestry, for example. Any disease that carries a genetic risk—that risk is then revealed. I go back to a comment that Professor Morris said earlier about how we manage the risk and those five safes—the principles of constructing trusted research environments where we can bring data together and manage the risk appropriately in a transparent fashion. UK Biobank is a stand-out example of where genetics, linked in with other information, have provided huge scientific advances in the understanding of the genetic contribution to human disease. We should think about the environments where we can manage risk in a trustworthy and transparent fashion. The use of trusted research environments in particular, and the five safes—such as safe people, safe projects and safe outputs—is the way to manage the risk that comes from genetics.

Q40 **Katherine Fletcher:** So you are saying that it is basically impossible to anonymise an individual's profile? However, it is possible to make sure that the people who could trace that back to an individual are within a trusted framework.

***Professor Holmes:*** There are different ways of anonymisation. It is possible to fully anonymise data, using a technique called differential privacy. However, that blurs the data so much that it becomes practically impossible to learn or benefit from that data. Imagine taking pictures of the Committee today; if you wanted to anonymise it you would start to blur the images. To give statistical guarantees that the data remains anonymous, you would have to blur the information—blur the pictures—so much that there would be no distinguishing features that would allow you to link those to clinical outcomes and learn. The best output is not to anonymise the data—use it in its most granular form—but link it in trusted research environments where we can manage the risk in a trusted and transparent fashion.

Q41 **Katherine Fletcher:** I have a final question. My parents participate in Biobank, so while I have never given my DNA to a study, effectively it is all there under the custodianship of the brilliant people who do great research. Mr Booth was talking in the previous panel about consent; my parents have fully given their consent, but I haven't. However, all of me is in those samples. Do you have any perspective on what we could do to take the public with us, so we can benefit from the enormous possibilities for good that this presents?

***Professor Holmes:*** As was said in the session before us, it is absolutely essential to bring the public on this journey and communicate to them the benefits and the risks. We need to be realistic about the risks that come from sharing health data. We particularly need to communicate the actions that are taken when breaches occur. We know that things will go wrong—things always go wrong. How do we set up a system that can manage that when those things happen? Communication of risk is not my particular expertise.

Q42 **Katherine Fletcher:** Let me get Dr Lewis-Brown's perspective.

***Dr Lewis-Brown:*** Thank you. There is not that much more that I would add to the discussion that you and Professor Holmes have just had, other than to emphasise that it is incredibly difficult to fully anonymise data. As Professor Holmes has outlined, there are ways, but we then lose some of the value of that data. That is why it is so important to have a secure and trustworthy framework as a safety net to support those shortcomings. That safety net, as Professor Morris described earlier, is based on the five safes: safe people, safe projects, safe places, safe outputs and safe data. It is worth emphasising that, although it may not be possible to fully anonymise some types of data, it is certainly possible to anonymise the outputs of that data, so that the only information that is released as a result of the analysis is fully anonymised and there is no risk—or a very low risk—of reidentification.

You talked about whether our DNA sequences can ever be fully anonymised. I think it is worth pointing out that while whole genome sequencing plays a role in research, there are other sequencing technologies and approaches that are just as important. In cancer, we rely on panel testing, where we look for specific cancer genes that are present in patients or tumours. By doing that, we do not have to get a view of the whole genome; we are just looking for certain parts of that genome that we know are closely related in a causal way to cancer.

Q43 **Katherine Fletcher:** That is an extremely helpful point. While you cannot anonymise the DNA sequence, you can perhaps protect the rights of the individuals for almost ancillary losses. Say, for example, that I were carrying the genes for a severe peanut allergy, and someone was very cross about what I said about Ukraine. There is the potential for those two things to be weaponised. I do not say it glibly. Personalising medicine in the 21st century can very easily be followed by personalised warfare. Could you only display part of the data for the public good and leave the bit of data saying that I have a severe peanut allergy somewhere else? It

is just an example; I do not have an allergy. I am trying to granulise it.

***Dr Lewis-Brown:*** Yes.

**Katherine Fletcher:** Brilliant.

Q44 **Aaron Bell:** I want to go back to the fragmented data ecosystem we have in this country and the number of different bodies that have an interest in data, as was referred to in the first panel. In what ways does the infrastructure we have hinder or support your research, innovation and efficiencies?

***Dr Lewis-Brown:*** The current UK data infrastructure does hinder and support research and innovation for cancer research specifically. In terms of its support, we do not take for granted that the UK data infrastructure provides patient data for research purposes. Not every country in the world has that. It does that through a system of trusted research environments, which is an approach that Cancer Research UK absolutely endorses. We have an almost entirely digital primary care system. Pretty much all GP consultations are recorded electronically these days. We have some standardised ways of recording patient data. That standardisation is really important when it comes to research, for example with the use of ICD-10 and SNOMED CT codes to categorise patient data. There are some datasets that are of exceptional quality, such as cancer registration data in England. There are many ways in which the UK data infrastructure supports research and innovation. We are very grateful for that, as are the cancer patients who we serve.

That is balanced by a number of ways in which the system does hinder and hold back the research we could be doing. At the top of the list for Cancer Research UK is timely access to data. Our researchers are typically waiting not weeks or months but years to get access to data. I mentioned high-quality cancer registration data. It takes around two years for that data to be curated to the exceptionally high quality that it is curated to. Once it becomes available for putting in an access request, it can take around a year to get access. By the time our researchers get hold of that data for their analyses and cancer research projects, it can be three years old. We feel that that is not acceptable. We need to find ways to speed up timely access to data.

The question might come up later, but there is a particular concern about the merger of NHS Digital into NHS England and that that could cause further delays, but perhaps we might pick up on that later. I have a couple of other points to make here. There is data that we can't get hold of. In the UK, we have some fabulous cancer screening programmes, for example—breast, cervical, bowel—but there is no route to access that data for our researchers, and the fragmentation and lack of standardisation in data collection is a problem. We have multiple trusted research environments, which is not a problem in and of itself, but they do not necessarily talk to each other as well as they could in terms of interoperability, and there is no accreditation process to make sure they are standardised. There is quite a lot of inconsistency across the four

nations. I suppose a point to make as well is that much of secondary care data is still paper-based, and we obviously need to move on from that position.

Q45 **Aaron Bell:** You said in your written submission to the Committee that you need better digital skills in terms of recording the data in the first place. Presumably, if data were recorded better in the first instance, that would also reduce the time it takes to curate it.

***Dr Lewis-Brown:*** Yes, absolutely.

Q46 **Aaron Bell:** At the end of the last panel, Professor Morris expressed some concerns about the representativeness of data given opt-outs. Is that of concern to Cancer Research UK?

***Dr Lewis-Brown:*** Yes, one of the benefits and advantages of using health data that is held by national data custodians is that there should be a pretty good representation there because such a high number of the population have records in our NHS. The representation should be very good, so we are very concerned about the opt-outs and the granularity of those opt-outs. What communities have the highest rates of opt-out, and are those people from groups that are under-represented in research? If that is the case, that undermines the representativeness of the data within the NHS system that is critical to cancer research.

Q47 **Aaron Bell:** Before I turn to Professor Holmes, is there a single suggestion you would make to simplify the ecosystem, or is it more complicated than that?

***Dr Lewis-Brown:*** It is always more complicated than that, but if we were forced into picking one, at the top of our list would be streamlining access to data for accredited researchers, done in a way that channels meaningful involvement with patients and the public. I do not think it is really possible to separate those two.

Q48 **Aaron Bell:** Thank you. Professor Holmes, I have similar questions for you. In which ways have you found that our infrastructure, our ecosystem, either hinders or supports what you do at the Alan Turing Institute?

***Professor Holmes:*** First up, there is huge potential in the UK to bring data together through the national health service, so it is really a question of moving towards interoperability. How do we open up the taps that will allow data to flow, which is a governance issue, and then once those taps are open and the data is in a trusted research environment, what are the challenges to drawing meaningful inferences from data that is brought together?

As I said, there is huge potential in the UK. As for the things that hinder it, first of all, the governance structures are quite local and fragmented. We at the Alan Turing Institute have experience of trying to bring together data from two major NHS hospital trusts, involving some of the leading data scientists and two of the leading hospital trusts in the UK. It is a challenge—with the best will in the world, it is a challenge. How we

streamline the governance structure while keeping it transparent and according to best practice, as Professor Morris indicated, is one big challenge.

The other challenge is, once you have the data governance in place that allows the data to flow into, say, a trusted research environment, what are the data standards that allow us to knit and weave that data together—that allow us to apply algorithms in a robust and reproducible fashion? That comes down to data standards, so those are the two things that would probably most hinder algorithm design and algorithm use from our perspective at the Alan Turing Institute.

Q49 **Aaron Bell:** You have both mentioned trusted research environments already. Clearly, you are both supportive of those environments, but how well do you think they are working? Are they being properly funded and supported? Are they a bit fragmented across the landscape of our current ecosystem? Professor Holmes?

***Professor Holmes:*** We are in the fairly early days of trusted research environments, but there are some stand-out examples, in particular the ONS, and in Wales SAIL, which is a trusted research environment for health and social data. We are learning through doing, but from our perspective at the Alan Turing Institute, of all the different ways of bringing data together, the trusted research environments seem the most appropriate for bringing data together with those safe principles that we have heard mentioned multiple times today.

Q50 **Aaron Bell:** Some of the evidence we had from Cancer Research and from HDRUK was that TRUs are not necessarily working together yet, and they are all setting up their own standards and structures without the interoperability you have just called for. Are you concerned about that, Professor Holmes?

***Professor Holmes:*** Again, it comes down to interoperability. As statisticians, computer scientists and mathematicians working on the algorithms, what we need to get the most benefit is interoperability across trusted research environments. We have institutes, particularly HDRUK, that are working to make those connections across the trusted research environments to bring that notion of interoperability to data standards. Certainly, supporting that initiative is hugely beneficial to getting where we want to be.

Q51 **Aaron Bell:** Dr Lewis-Brown, what is your perspective on the way TREs have developed so far and whether there is a functioning network of TREs yet?

***Dr Lewis-Brown:*** As Professor Holmes says, it is early days, but it feels like it is on a positive trajectory. There is a need to have some kind of central accreditation system for trusted research environments, so that we can make sure we are all working towards the same standards and towards interoperability, because that is where the real value of trusted research environments lies.

We warmly welcome the Government's recent announcement that they will invest up to £200 million that will go towards improving access to NHS data through trusted research environments and digital clinical trial services. We hope that there is a sufficient level of public and patient involvement and engagement in those plans. It would be good to see how that funding will be targeted to really realise the potential of this network of trusted research environments.

I think there is a degree of inconsistency. The data access pathways and the data flows are working better in some parts of the four nations than others, and it feels as if there are shared learnings that could help to drive the whole system forward. As you know, we support the trusted research environment framework and its use, and we look forward to there being some kind of kitemark system, an accreditation system and harmonisation mechanisms.

Q52 **Aaron Bell:** Presumably that needs to happen now, because as it is growing is the best moment to set those standards.

***Dr Lewis-Brown:*** Absolutely.

Q53 **Aaron Bell:** Briefly, you indicated that you wanted to comment on the merger of NHS Digital and NHS England, so let me give you a chance to do that now.

***Dr Lewis-Brown:*** Thank you. Our main concern about the merger that is under way, with NHS Digital moving into NHS England, is that there may be further delays to data access for our researchers. We know from experience that when datasets that were held by Public Health England were moved into NHS Digital, the data access process was paused for two months, which led to a backlog. I don't think we fully understand what the size of that backlog is, but it has definitely impacted on data access waiting times.

As I mentioned, for the kind of research we do, which is life-saving cancer research developing new treatments, our patients cannot afford unnecessary delays as a result of mergers. I guess we would encourage this inquiry and Government more broadly to prioritise the speed at which cancer researchers, and health researchers more generally, can get access to this data for their life-saving research.

Q54 **Aaron Bell:** I can see speed is the issue. Do you have any ethical concerns about the merger? In the first panel, Phil Booth said that we would have the situation where NHS England can commission the data that it wants without necessarily going through the same process of consent. There would be consent, but he thinks NHS England becomes a customer rather than a regulator.

***Dr Lewis-Brown:*** I think it is critically important. What we hear from our patient networks is that no one can get special treatment; we all need to abide by the same rules and be managed in the same way when we are applying for data access, and that is no different for NHS England or any other body.

Q55 **Dehenna Davison:** I should declare an interest; I have been a doner to Cancer Research UK for a number of years, so I will put that on the record. Most of what I was going to cover was actually covered by Katherine's questions on the privacy concerns around genomic data, so I will touch on privacy-enhancing technologies and the sort of scope that they have for protecting privacy while ensuring that the data available to researchers is as full as possible. Professor Holmes, can you give us an overview of some of the PETs that are available and how useful you think they are in this context?

***Professor Holmes:*** I would distinguish between three levels of privacy-enhancing technologies. The one that we have heard a lot about is the use of trusted research environments and the safe mechanism—safe people, safe projects, and so on—set up around it. At that level, that allows data to come in its rawest form, as the original data, and there are safeguards in place on transparency, the trustworthy use of that data and what the data then feeds into.

At the next level up, there are things such as federated learning, where you do not allow the data to flow into a single place; you might keep the data separate, perhaps because you have an international collaboration and the UK does not want to transport its data to Denmark, and vice versa. That offers a level of data protection and privacy because the data does not move. It could sit in trusted research environments at the host institutes. As you move up that level of privacy, you lose the ability to learn more, because the algorithms then have to be moved to the data; you do not have the opportunity to visualise the data and work closely with it.

At a higher level still, you have things around differential privacy, which I mentioned earlier, where you just release the data—you do a public release of the data—but that data is fully anonymised. You have to be very careful when you do that because we can triangulate with data. We might release a dataset that we feel is anonymised— in the example that I gave, we might have blurred our images—but when you combine those blurred images with other blurred data, you can quickly triangulate and identify individuals. We know mathematically how to correct for that, but you have to add so much blurring that the data becomes pretty useless for inference.

To recap, there are three levels. As trusted researchers wishing to push science for public good, the lowest level—the trusted research environment—is the most promising of those levels for doing good science and building accurate clinical decision support systems.

Q56 **Dehenna Davison:** Do you think that that level, if it was universally adopted, would give the public, who can be very concerned about their own data privacy, enough reassurance that the right measures are in place?

***Professor Holmes:*** It is about bringing the public with you by having public and patient involvement right from the go in the design of the

systems and of the safeguards, understanding what science, analysis and algorithms are being applied, and communicating back the positive aspects of the outcomes from using the public data for the public good. It is about management and communication.

That becomes particularly important with opt-outs, which we have heard a lot about. We were worried about opt-outs because they lead to unrepresentative data. Different communities, we know, have different op-out rates. We need to be proactive in the communication to those communities about the benefits of opt-in, and how we use their data for the public good. I know that Genomics England have been thinking carefully about that in a diversity programme about how they can start to address those concerns through communication and outreach.

Q57 **Dehenna Davison:** Thank you, Professor Holmes. Dr Lewis-Brown, is there anything that you would like to add to that?

***Dr Lewis-Brown:*** Not really. I agree with Professor Holmes's perspectives. As for whether that answer would reassure the public and patients, we absolutely must ask them, and what we are hearing from our own patient involvement network is that there is a great appetite to be involved in these discussions. They want to be involved in these discussions. They want to be educated and made well aware of all the risks involved and to talk very openly and transparently about those risks. They want to be listened to. They want to know that the risks that they perceive or risks that they outline, whether real or perceived, are addressed and mitigated, and those mitigation measures must be well communicated. Some of these privacy-enhancing technologies are complex, so we need to think about how we can communicate them to the public so that they can make informed decisions about whether they feel comfortable with data sharing when these sorts of technology are in place.

Q58 **Dehenna Davison:** Do you think that the onus for that communication should be on individual bodies that are actually using the data, or do you think that more should be coming directly from Government?

***Dr Lewis-Brown:*** It definitely needs to be both. Anyone working with health data for research purposes, or any other purposes for that matter, has an obligation to communicate the work that they are doing to the public and also to involve members of the public in their work. So I think the onus is on us all.

**Chair:** Thank you very much indeed, Dehenna. I thank both our witnesses, Professor Holmes and Dr Lewis-Brown.

## Examination of witnesses

Witnesses: Professor Aziz Sheikh and Professor Chris Molloy.

Q59 **Chair:** We are now going to go to our final panel of witnesses. We have in person Professor Chris Molloy, chief executive of the Medicines Discovery

Catapult, which helps to translate discoveries into applications that can treat patients. Joining us virtually, we have Professor Aziz Sheikh, director lead for digitally enabled trials in the BREATHE Health Data Research Hub, which, as the name implies, is focused on respiratory data and advances in medical science in that area. Professor Sheikh is also director of the Usher Institute and dean of data at the University of Edinburgh. Thank you both very much for joining us.

Professor Molloy, I will put the first question to you, given the work that you do in the catapult. Some of the evidence that we have had is from pure research settings. You are involved in the translation of research into practice. How important is data analysis in your work?

***Professor Molloy:*** Thank you, Chair. I should say that I also chair the board of directors of two companies in the UK that use real-world evidence data to improve the quality and speed of clinical trials. They are Exploristics Ltd and NorthWest EHealth Ltd.

Within the catapult itself, everyone recognises that medicine is a data science, and the practice of healthcare is as much a data science as it is a physical science. It is about data-driven decisions; and without high-quality data, you make poor-quality decisions, whether that is about what new drug targets to chase, any evidence for whether a diagnostic indicates a disease is progressing or not, or how you collaborate with one another. Getting the data right is the first foundation of good drug discovery, medicines discovery, diagnostics discovery, and any new intervention.

While we have in the UK an enormous opportunity, which I think all my colleagues have spoken to and I absolutely acknowledge, we have also created an absolute briar fence of a barrier, particularly for SMEs to come and access those data. The assumption is the data do not flow. The assumption is it is hard. The assumption is it takes time—too long.. According to a report that we published with the ABPI in 2019, even in the best case of access to data, fewer than 45% of data attempts or attempts to get research data through from the existing systems are successful, and that is the high water mark; the lowest is about 20%. The assumption now in the industry is that these are data vaults—inaccessible and museum-like. They are useful to researchers in academia, but with the timescales for getting hold of them and with the constant increase in very well meaning regulation, we are making these data inaccessible. In a global market for ideas, invention, companies and investment, companies are therefore pushed to other nations where the data are secure, but where the access to those data is measured in weeks and months, not months and years. If we are to enable our communities, we must get better than that and assume data sharing as a default.

**Chair:** Thank you, that is very clear.

Q60 **Chris Clarkson:** I have a question for both panellists that I have posed to every other panel, and I will start with Professor Sheikh. To what extent do the Government's data sharing proposals address the complexity of data sharing at the moment, and what are the dangers of their proposals

to streamline the way that data is dealt with?

***Professor Sheikh:*** Thanks for having me. I largely concur with the earlier witnesses. At the moment, the process is very complex and pretty opaque, even for people who are long in the tooth and have been working for a long time with the fantastic data assets we have in the UK.

I am not sure that additional legislation is needed, but we need to get to a position where the expectation is that the data that the NHS is generating, as well as the data that we have, are being used and repeatedly reused in safe and trustworthy ways for public benefit. So much of this is about helping navigate the information governance landscape, as earlier speakers have also mentioned.

**Chris Clarkson:** Thank you. Professor Molloy, I saw you nodding there.

***Professor Molloy:*** Navigation is everything. We do not need more legislation in this area. The information governance that we have for health data is over and above GDPR. Simplification will not lead to a decrease in privacy. Simplification should lead to an increase in use, and that is how we should measure this, not by the input but by the use of the data that goes through that system. That is the metric.

Q61 **Chris Clarkson:** Fantastic, thank you. I will also pose the question that I asked Dr Lewis-Brown. How might the Government's plans impact on EU adequacy agreements and data? She made it abundantly clear that for niche datasets you have to go out of the country, and that consistency in data is quite important.

***Professor Molloy:*** She is better placed than I to see the impact of some of those things. There is a theoretical risk there. I don't believe that our data are insecure or that our access to data is insecure. We may have made it difficult, but we have certainly made it secure. I don't believe that the proposals that are being made here should give significant cause for international concern that we have somehow just opened the doors wide. What we are doing here, which I believe is necessary, is to simplify access to consented data.

Q62 **Chris Clarkson:** Looking specifically at your role, when you talk about dealing with innovative SMEs, how much does compliance with existing regulations inhibit the ability to innovate?

***Professor Molloy:*** Enormously, because we have driven it to the assumption that UK data is almost impossible to get hold of for an SME. If you think of most UK biotechs, 60% of them have five people or fewer. Their funding is over a short time horizon, often a shorter time horizon than it will take to go through the process of accessing the data. If you can't access the data, then you can't get your investment because you can't validate your hypothesis, so you are in this Kafkaesque approach to the problem. Inevitably this drives innovators to go to places where they can get that validation data consented and available in less than geological time. It is really important.

I don't take anything away from the concerns of patients that they don't want to see their health data on the internet—I don't want to see that. Neither do I want to see my banking data on the internet, but there are plenty of controls to make sure that my banking data doesn't get made available on the internet, and there are even greater controls around health data today. SMEs and larger companies in the UK need to be both enabled, maybe by pre-accreditation, so that they can fast-track through some of those early processes, because every new request starts at the beginning again, with a new hypothesis, a new justification and so on and so forth. Even if that same company has done this a hundred times, they start again at first base. It just takes time, because we are not accrediting people for the basics and then allowing them, in effect, to start the process at second or third base and get to the end point in a sensible amount of time. That will save us time ultimately.

Q63 **Chris Clarkson:** Have we learnt any lessons from the experience of covid-19? I pose that question to both witnesses.

***Professor Molloy:*** Covid-19 showed us an exceptional collaboration between the NHS, academia and industry. Leave one of those people out of the room and this will fail, and it should fail. Industry needs to be in the room, not just as an enabler and consumer of this, but as a co-developer of this regime and system that we have. This is not something about which we should say: "We do all this in a private sector environment, then industry is somehow the rapacious consumer of the data at the end." This has to be a co-development. That is what we need to learn from covid-19, and if we forget that, we will go back to the time when we said: "Here's health data", and, "Here's academia", which is deemed to be research, but then there are these other people, who are deemed to be the commercial exploiters of this information. That is an extremely dangerous set of language that can be used, and a dangerous assumption that almost all documents in this area make, where industry is sort of airbrushed out, so that research becomes an academic endeavour, which it most certainly it is, but without the industrial rigour, you do not deliver products to market and you do not deliver new products and services to patients.

Q64 **Chris Clarkson:** So it is understanding the application as well as the science.

***Professor Molloy:*** Absolutely so.

Q65 **Chris Clarkson:** Professor Sheikh, is there anything you would like to add?

***Professor Sheikh:*** Covid-19 has shown us what is possible, for example, having data provisioned as close to real time as possible to inform all aspects of decision making—data often being provisioned daily. We are seeing those data being relayed and used by chief scientific advisers and chief medical officers across the UK to make really important decisions.

Your earlier witnesses spoke about how in other contexts sometimes, there are delays of months or even years. In my own experience, often, 80% or 90% of time on a project is spent dealing with information

governance and release considerations, so I think that somehow we need to learn the lessons of covid-19. We are not through covid-19, by any stretch, and we need to move pandemic preparedness recovery, but there are other "pandemics"—in inverted commas—that we are dealing with: cancer, diabetes, mental health, which have all been mentioned. Covid-19 has shown us what is possible. It would be an act of absolute disaster if we were to revert to our models pre-covid.

Q66 **Chris Clarkson:** Thank you. I will stick with the theme of regulation, because I asked the other panels about the regulation of AI, which will play an increasing role in research. First, Professor Molloy, what are your thoughts on the adequacy of existing regulation of how AI is used in research? What potential gaps are there, and what pitfalls do you see?

***Professor Molloy:*** AI is a technique; it is not a thing in its own right. It is the use of an algorithm, which can be anything from a simple mathematical calculation upwards. It is unhelpful to talk about AI as though it were one single thing that we know and that we can put in a box. Clearly, the use of algorithmic tools working on the data will give us new products, services and techniques. If those are used in clinical practice, as drugs and devices are, they have the potential to be trialled alongside existing approaches.

There is a perfectly reasonable and sensible way to regulate that, which is to say: "Does this make better decisions than the ones used by clinicians?" Clinicians, in themselves, are biased inevitably through their own experience. They are machine-learning tools; they learn, and they are not stuck in aspic. Like all clinicians, they are taught at medical school; when we go to our doctor, we do not ask, "May I see the textbook that you learnt from 25 years ago, just to make sure that you are up to snuff?" They then learn through experience. All we can do, at the beginning of regulation, as we do with drug therapies, is ask, "Is this fit for purpose? Does it routinely and robustly do this thing well, or either as well as or better than existing current practice?" Then it will learn. We need to constantly, as with drugs, do in-market trialling to make sure that they are still meeting or exceeding expectations. That is the way to regulate that, as we do with other medical interventions.

Q67 **Chris Clarkson:** Would you agree with Professor Holmes on the last panel that it is more augmentative than a means in itself?

***Professor Molloy:*** Very much so.

Q68 **Chris Clarkson:** Professor Sheikh, is there anything you would like to add?

***Professor Sheikh:*** I largely concur with what has been described there. One other point to make is that our capabilities of using data-driven approaches are enhancing phenomenally. That is because the datasets have really developed during course of the pandemic. We have fantastic computational abilities, and we have an increasing cadre of really skilled analysts in the UK who are trained in working with these large datasets. Our ability to work with these datasets is going to increase very

substantially in the coming years. The regulatory framework we need must recognise that this is a continuous learning process. It needs to go through derivation and validation phases. Ultimately, the proof of the pudding is in whether it is improving outcomes or has the potential to do so. I think we need to look at it through that lens. If we play our cards right here, we can position the UK as a real world leader in this respect.

Q69 **Chris Clarkson:** Would you say it is probably better to approach this in terms of guidance and regulation, as opposed to additional prescriptive guidance through legislation? It is about ensuring that the regulation of this is as flexible as possible to keep pace with the fact that this will be a changing, developing system.

***Professor Sheikh:*** I don't think legislation is the answer here. I think the answer is a regulatory framework that is as enabling as possible, while obviously looking at safeguarding issues out there. It must also be kept under review, because this field is moving at pace.

***Professor Molloy:*** It is about being enabling and agile in the regulatory space, because this will move faster than most regulatory systems can keep up with. We already know that around the world most regulatory systems are catching up with digital health. It will be a constant effort for regulators and providers to work together and co-develop on this. That is not to suggest that it is poacher turned gamekeeper, but it is a co-development of these regulations. As new technologies come forward, the regulators need to understand that and work out how best to regulate it. Unless they are working in lockstep, it becomes adversarial, which does not help anybody.

Q70 **Aaron Bell:** Professor Molloy, I would like to follow up on what you said to Chris Clarkson about the obvious benefits of private sector companies being able to get access to data. Clearly, there is a lot of evidence out there to show that the public have much lower levels of trust in sharing with commercial organisations. What solutions could you offer to try to mend that trust? Is it about drawing some red lines around the purposes the data could be used for?

***Professor Molloy:*** Patient and citizen engagement is certainly important. We live in a sort of Beveridge 2.0 time, when, instead of having health services done by the establishment to patients and citizens, we are moving into an area where citizen engagement with their own health is so vital. It is not just about people taking more care of their health, but about engaging with healthcare providers, either in providing data or biosamples or getting access to Fitbits and other things to enable the health service to help them. I see this as a really important conversation.

Nobody wants people to opt out, but if people feel strongly enough that they want to opt out, we should let them. We have things like the NHS app, which covid has actually driven the adoption of. That might be a platform on which to engage with patients and provide them with a list of options that they will be able to swipe left or right on. That might be a way

in which we can get a more two-way, agile and high-fidelity touch with patients.

We also have to look at what that gives back. We can't just say that something good might happen in 30 years' time. We have the option as a nation to look at the financial returns that can be made, and if data builds a new cancer ward, like the national lottery has built great sets of infrastructure around the nation, that engages patients with the power of their data: "It has built a cancer ward for my grandmother. I should give my data and I should allow those data to be used."

The role of medical research charities is really important. I don't believe that they necessarily need to have the burden of being the arbiters of whether data gets provided or not, but they certainly are a very strong voice.

Q71 **Aaron Bell:** What should be the quid pro quo for private companies getting that data? Should it be financial or should it be an obligation to share the results of their work with the data?

***Professor Molloy:*** There is a range of models, which a PwC report prior to the pandemic reviewed. We can look at intellectual property, and we can look at long-term returns to a sovereign health fund or something centralised that would enable the critical mass of money to be built up. That would help people to recognise that this is a real thing, rather than just necessarily a local thing. We can put back to citizens the products and services that have been built using their data and that are now helping other patients. People are generally quite altruistic in this regard—it helps others. Covid has been a big "helping others" time, and I think we can use that momentum.

Q72 **Aaron Bell:** You were here for Phil Booth's evidence at the beginning. He said that unless we get this right, there is going to be a ratchet effect of people opting out. You have just said that you respect people's right to opt out of this sort of stuff. Are you not concerned that unless we get this right soon, you are going to have a really big ratchet effect and half the population are going to opt out of this sort of stuff?

***Professor Molloy:*** There is always the potential for that. I think that people are more altruistic than that, and I would not foresee a time when half the population might choose to opt out if the conversation is performed sensibly. Work was done by organisations like Hopkins Van Mil just before the pandemic; there were deliberative engagements with citizens, and citizens do want their data to be used for good, as long as they are well governed and there is some form of fair return. In general, that is what happens; whether it is patient groups or the general population of well people, folks are generally comfortable to do that. There is always the potential for the headline-grabbing, "My data's been sold to—", and I think that that is extremely unhelpful in what should be a balanced discussion about the use of data and its return. I do not see biotech companies or other companies having a specific interest in

identifying the individual medical data and making that public. There is no incentive for them to do that.

Q73 **Aaron Bell:** Understood. Briefly, because we are running out of time, I asked the other panels about the fragmented data ecosystem that we have in the UK and the variety of people who have an interest in it. May I ask you both, starting with Professor Sheikh, what one suggestion for change you would make to simplify the data ecosystem in the UK?

***Professor Sheikh:*** I agree: I think it is very complex to navigate at the moment, particularly for those who do not use health data all the time and for those outside the UK who want to use data. I think what would help would be a single register of what datasets are potentially available. I know that Health Data Research is trying to do that through the innovation gateway. There is that model, or another model could be made available, but there should be a single register of what is potentially available, because we have a wealth of data in that respect, and then it is about helping to streamline the governance process so that we are able to respond to requests to access data within days rather than the months and sometimes years that it is currently taking.

Q74 **Aaron Bell:** Thank you. Professor Molloy?

***Professor Molloy:*** Streamlining? Yes. I would like to see the pre-accreditation, which would get people, and the ability of citizens to opt-in—with a sensible graduation of opting-in to primary and secondary uses, and so on, in a really simplistic way, but with an assumption that actually opting-in is good and helps others. That requires public and political discussions with people about what is good for them and good for others.

Q75 **Katherine Fletcher:** Thank you, both, for your helpful insights. There are two separate topics that I wanted to come back to. First, in this Committee, we like to try to make tangible recommendations. I have heard, all through this morning, people saying "Yes, it takes ages to get data out when you make a research request." There are people in that system; what is incentivising them to be so slow? I am sure that, if they are working in that field, they do not believe that. Are they worried about personal risk because they have a lack of clarity? I can see nodding, so I will start with Professor Molloy, then I will bring you in, Professor Sheikh.

***Professor Molloy:*** Let me start by saying that I do not think anyone is sitting there, trying to stop things from happening. Clearly, there are very well-meaning, sensible regulations about ensuring that data are protected. However, that sometimes leads to the default answer being "Well, why should you want it?", rather than, "Yes, but let's just check that the protections are there."

It is perhaps an old, legacy view that, "That the data are mine," on an institutional, or even national, basis, and that the assumption is, "You need an extremely good reason why," rather than making the suggestion, which I know is in some of the Government proposals, that data sharing is the default, and then you work back from there. This has come by the building up of regulation and legislation over many years, which has

caused almost an assumption that you need an absolutely gold-standard, copper-bottomed reason—and other mixed metaphors—to get hold of data, rather than the assumption being, "You should get hold of it, but let's make sure the safeguards are there."

Q76 **Katherine Fletcher:** Do you concur, Professor Sheikh, that it is effectively a system that is regulating to prevent illegitimate access, as opposed to preventing positive, secure access. Is that your experience of it?

***Professor Sheikh:*** Yes, I would concur. Quite often, the easier decision to make, from a data controller's perspective, is that it is safest not to share. We need to really move away from that position to one where the default is that the expectation is to share.

There are other dimensions to this too. We quite often look at this from an academic perspective; individual research teams will think of a dataset that has been generated as their own, proprietary interest. As datasets are funded, and research is funded, the expectation, again, should be that the default position is data sharing. As we are developing our ethics applications, for example, again, consent models should be really working in data-reuse considerations right from the outset. The whole system, at the moment, is on the back foot, and we need to get it on the front foot.

Q77 **Katherine Fletcher:** That is a helpful segue into my second question. I am very struck by a lot of the British public's concerns about these big businesses coming in and taking their data. It is a well-grounded sense—which I agree with—of, "I'm not having these people making a profit and getting something for nothing." A way to bridge that is the idea of an altruistic act with something that they can see, as a beneficiary.

I am very struck by the idea of a sovereign health fund. Professor Sheikh, I will come to you first, because you are obviously dealing with this day-to-day. If we could incentivise the British public to give up their data—with some safeguards on some known issues—because it will not only benefit outcomes but help pay for the NHS in the future, with all of the commercial revenues from the IPO, or whatever, is that something that you could work with to achieve your aims?

***Professor Sheikh:*** I am very supportive of that suggestion. If we take a 10 or 15-year time horizon, the majority of the data that we will be talking about are going to be generated by citizens themselves, working through apps and so on. NHS data are going to be a small proportion of the data ecosystem. Industry is fundamental to this. You have heard at length about the five safes. A lot of this can be regulated—what projects, what people, what kind of outputs and what uses. Some way of sharing the benefits when industry is accessing data is only fair and right, and it is something that the British public would get behind.

Q78 **Katherine Fletcher:** Thank you very much, Professor Sheikh. If we could find a model where people pay to subscribe to the data to get those revenues to our NHS or health services early, as well as a benefit over 10 to 15 years as ideas turn to proven concepts turn to commercial, would

you support that?

***Professor Sheikh:*** Yes.

**Katherine Fletcher:** Cracking.

**Chair:** Thank was the most concise answer of the day, but it was very clear. I thank Professor Molloy and Professor Sheikh, and all our witnesses, who have got our inquiry off to a cracking start today. We are grateful. That concludes this meeting of the Science and Technology Committee.