HOUSE OF COMMONS

Science and Technology Committee

Oral evidence: The right to privacy: digital data, HC 97

Wednesday 11 May 2022

Ordered by the House of Commons to be published on 11 May 2022.

Watch the meeting

Members present: Greg Clark (Chair); Aaron Bell; Katherine Fletcher; Rebecca Long Bailey; Carol Monaghan; Zarah Sultana.

Questions 180-272

Witnesses

I: Mark Sowden, Government Chief Data Steward, Government Statistician and Chief Executive, Statistics New Zealand.

II: Christian Reimsbach-Kounatze, Information Economist and Policy Analyst, Organisation for Economic Co-operation and Development.

III: Professor Ben Goldacre MBE, Professor of Evidence-based Medicine and Director of the Bennett Institute for Applied Data Science, University of Oxford.

Examination of witness

Witness: Mark Sowden.

Q180 **Chair:** The Science and Technology Committee is in session. This morning, we continue our inquiry into digital data and, in particular, its privacy aspects. We are very pleased to welcome Mark Sowden, who, geographically, is the furthest witness from Westminster who we have had in this Parliament. He is the Chief Data Steward of New Zealand, Government statistician and chief executive of the New Zealand statistics authority. Thank you very much for joining us, Mr Sowden.

Perhaps you could give us a little bit of background on the statistics regime in New Zealand. The Committee has heard about the great opportunities, particularly in the medical and health fields, for medical records to be used for advances in science. Obviously, that gives rise to important privacy concerns. How has New Zealand approached those dilemmas?

Mark Sowden: Sure. I will start with a couple of pieces of context. First, as you said, I play the role of Government Chief Data Steward, and part of that role is to get all the different arms of Government that work with data to work better and more effectively together. That is partly designed to make data more shareable. Part of what we are trying to do is raise the quality of data across the system so that data collected in one part of the system—be it the Ministry of Health or our Ministry of Social Development, for example—is of the same quality that we or other data-specialist agencies would collect. Part of how we address the data quality and shareability issues is by raising those common data standards. That is one piece of context.

Another piece of context is that we have a long seven or eight-year history of a thing called the integrated data infrastructure, or IDI, which is a data-linking infrastructure that we host in which different datasets from various arms of Government—I can go into detail if you want—including the Ministry of Health. Our data are held together and linked so that you can understand data about an individual person. That data is anonymised so that you cannot identify that it belongs to Mark Sowden, but you can follow a person who looks like me through the system, so you can look at different health outcomes, at different life course outcomes, and over time you develop a longitudinal view of different outcomes for different demographic groups of different people with different characteristics across society. Again, I am happy to talk about what safeguards and things we have put around that. That is part of the infrastructure.

For our own statistics work, we also have the Statistics Act, which is soon to be superseded by the Data and Statistics Act, which is going through our Parliament at the moment. That gives us and other Government agencies a specific permission to share data for research and statistical purposes. It actually overrides our Privacy Act—a basic privacy Act to protect the privacy of New Zealanders' data—and allows an opt-out from

it. It effectively allows anonymised data—not identifiable data about individuals—to be collected and shared for statistical and research purposes.

I hope that gives you a bit of an overview. I am happy to go into any of those areas.

Q181 **Chair:** Thank you very much. My colleagues will have more detailed questions but, to stay at the broad level for the moment, before the passage of the new Act of Parliament, to what extent is medical and healthcare record data available for research purposes? Is it in small groups, or, more generally, are the health records of the population of New Zealand available to researchers?

Mark Sowden: It depends on what is added to the IDI, the integrated data infrastructure. We make decisions and we have certain criteria that we use to ascertain whether to add certain datasets to that integrated data infrastructure. So, we have criteria around that.

All the health research proposals in New Zealand that want to access data go through our tests. We have what we call a "five safes" test: safe people, safe research, safe location, safe output and safe data. It is designed to make sure that access is only for specific research purposes and that that research is somehow in the public good, so we look specifically at the research question. We also have physical security around access to the data, we check the data again, and we go through the final research output to make sure that it is anonymised and that you cannot re-identify individuals from it. So, we have that test.

Also, all health research in New Zealand needs to go through a health research ethics panel, so that there is an ethical component as well. Only when those two tests are met can researchers access the IDI.

Q182 **Chair:** I see. On that first test, you referred to "we"—that is the New Zealand stats authority, is it?

Mark Sowden: Yes, that is correct. I as the Government Statistician have the authority to grant, or not grant, access to the IDI and what datasets we put in. So, there are the datasets that we put in, and I make those decisions about what datasets we put in based on data quality and ethics considerations, then I grant access for each individual research project to access that data.

Q183 **Chair:** Each project has to apply, in effect, to you, or at least you have a decision to make on each one?

Mark Sowden: That is correct, yes.

Q184 **Chair:** Briefly, tell us how the new Act will change that or parts of that? Does it disapply any aspect, or does it take it further?

Mark Sowden: No, all those controls will still apply. What the new Act does is to provide a more global power for the Government Statistician to compel people to provide data for research and statistical purposes. It is

not so much about who accesses it or what we put in that integrated data infrastructure; it is whether I can collect the data in the first place.

Chair: I see. So, it gives you powers to acquire data; it is not really about the safeguards.

Mark Sowden: That's right.

Q185 **Chair:** One final question from me before I turn to my colleague Aaron Bell. How would you describe the current state of public opinion in New Zealand on questions of privacy about this kind of data? Is it a high-profile issue? Is it very prominent and contentious, or is it a background issue that has not really caught fire, as it were?

Mark Sowden: By and large, it floats around in the background. It is not really a prominent issue. Because we are quite careful about it, you do not get headlines about data breaches and inappropriate use of data. What I would say, though, is that with covid-19 and all the use of data around understanding covid-19, actually the benefits of being able to use health data for research purposes has been brought to the fore, because we have done a lot of research and a lot of modelling, as all countries have, to be able to understand covid, the way it has flowed through the country and the epidemiology of it. We have worked hard to tell the story about how the sharing of and access to data have enabled us to do that. So, we have been a little opportunistic, I must admit, using that to tell the story of the benefit of data sharing to New Zealanders.

Chair: Thank you. That is very helpful. I will turn to my colleagues, starting with Aaron Bell and then Rebecca Long Bailey.

Q186 **Aaron Bell:** Your integrated data infrastructure sounds like what we have been referring to in our inquiry as a "trusted research environment". Is that basically it?

Mark Sowden: Yes, it is, pretty much. We have put a lot of safeguards around who can use the data and in what conditions, including the physical conditions in which you can use it. Effectively, you have either to pass some very rigorous security tests and remote access it, or to go into one of 30-odd data labs that there are across New Zealand that meet certain physical security requirements.

Q187 **Aaron Bell:** In our TREs—trusted research environments—researchers cannot take data out of the TRE; they can only export analysis results. Is that the same for you?

Mark Sowden: Yes, that's right.

Q188 **Aaron Bell:** Thank you. I have a few specifics, as the Chair alluded to earlier. How is an individual's consent obtained if their data is going to be shared with third parties, and particularly with commercial entities?

Mark Sowden: Generally—if we are talking about health data in particular—when you submit your health data, when you fill in the form, you are asked to tick a box that says your data can be used for research

and statistical purposes, so it identifies that you can. Then you are asked if it can be shared within Government and for research purposes, and then you are asked if you are happy to opt into sharing data for research purposes with more commercial entities. Within Government, it is an automatic right.

Q189 **Aaron Bell:** So that is an opt-in process. Approximately what percentage of people opt into the commercial research as well as the Government?

Mark Sowden: Most do. Very few do not, and I think it goes back to the previous question that was asked: there is quite a lot of publicity here about some of the benefits of health research. In addition to the covid-19 stuff I referred to, we have a longitudinal study of a group of New Zealanders, so every five years, there is a report produced about their health and life outcomes. That research is very carefully linked back to the data that is collected over time, and it is all part of promoting the constituency around willingness to share.

Q190 **Aaron Bell:** Once they have opted in or agreed to share, how easy is it for an individual to later find out if their data has been shared, with whom and for what purposes? Can they go back and make a subject access request, or whatever you call it there?

Mark Sowden: Yes. Under our Privacy Act, you would make a request, and any Government agency and any commercial agency that holds data on a New Zealander at that point is required to talk about how that data has been used, with whom it has been shared, and why.

Q191 **Aaron Bell:** Thank you. We have a national opt-out where people can stop their personal data being shared for research or planning purposes. I believe there is an opt-out for the HealthOne information system on the South Island, and I think you implied that Government data was basically opt-in for everybody and there was a separate opt-in for commercial.

Mark Sowden: Yes, that is right.

Q192 **Aaron Bell:** Is there a national opt-out where you can just say, "I don't want anything done with my personal data"?

Mark Sowden: No, there is actually not. Under the Privacy Act, New Zealand does not have opt-outs. What we have is controls about who can access the data, so that is our social contract, if you like, with New Zealanders: we have pretty rigorous controls about who can access the data, but we do not have the opt-out that you do.

Q193 **Aaron Bell:** With the opt-in to the commercial element that we alluded to earlier, you said "most". Do you have a number, or if you do not, would you be able to write to us with a percentage?

Mark Sowden: Yes, we can certainly find that out. I do not have that number off the top of my head, but I can certainly find that out for you.

Q194 **Aaron Bell:** Are those who have opted out able to opt back in? Is that easy?

Mark Sowden: That would be fairly unusual, and you would have to go quite out of your way to do it. You are not asked continually, "Do you want to opt back in?"; you would actually have to decide, "I want to contact that organisation that I know was using it", so it would be quite difficult and pretty rare.

Q195 **Aaron Bell:** One of our concerns from the earlier evidence we have heard is that there is a bit of a ratchet effect. We have a national opt-out, as discussed, and every time there is some sort of story about data being shared—maybe not inappropriately, but it can be presented as such—more people opt out, and we never get them back. We are also concerned about the opt-outs not being statistically accurate: they are disproportionately from ethnic minorities, perhaps. Do you have that problem as well—that there is a bias in the kinds of people who opt out, young people or ethnic minorities?

Mark Sowden: Yes. Our Māori—our indigenous people—tend to be quite highly represented in those who, I would say, at least have a suspicion of what commercial enterprises are doing with their data. Again, I do not have the figures in front of me, but I would imagine that they have a higher proportion of the opt-out. We have an issue around Māori people's participation in the data system more generally. I would say we probably have a bit of a digital divide, or a data divide if you like, between the comprehensiveness of the data we have on our Māori population versus that on our general population. That is a wider issue for us.

Q196 **Aaron Bell:** Again, would you be able to share percentage figures with us by a subsequent email or letter?

Mark Sowden: Yes, I am happy to share what we have, of course.

Q197 **Aaron Bell:** Finally from me, what engagement activities have taken place in New Zealand to improve the public's trust in the sharing of data, and which of those approaches have been successful or unsuccessful, in your opinion?

Mark Sowden: The approaches that have been successful are the ones that talk about the public benefit. Quite a lot of different research projects were done that concluded that New Zealanders are relatively happy for their data to be shared if they think that will benefit themselves or there is a clear public benefit from it. I have mentioned a couple of examples where, around covid and the longitudinal study, we publicise very heavily where we got this data and why it is useful. The aim is to encourage others to do it. Stats New Zealand also undertake regular trust and confidence surveys to understand about New Zealanders' confidence in data sharing and our handling of the data. Also, we do a five-yearly census—like you do—and every time before we do that five-yearly census, we undertake quite a big exercise: "What do we do with this data that we collect from New Zealanders? How do we share it? How is it used?" Again, those all contribute to that building of the public understanding and almost the public commitment to saying, "The use of my data is somehow in the public good."

Q198 **Aaron Bell:** You mentioned covid, which obviously was exceptional. Did you make the data more widely available? So much research was done on datasets that were made public by Governments. Did you have a dashboard approach? Did you have an API that people could get data from about not necessarily specific cases but case numbers?

Mark Sowden: Yes, we had both those things. We put out a covid-19 data portal, which Stats New Zealand posted on behalf of all of Government, and sitting underneath that was an API, where they could get down to the underlying data, but they had a dashboard where you could create your own graphs and charts and things. Again, if you are interested, I am very happy to share a link with the Committee if that would be useful.

Q199 **Aaron Bell:** That would be extremely interesting. Did you find that third parties and even enthusiastic amateurs produced research on that that hadn't been done by you or anybody else?

Mark Sowden: Yes, that's right, and part of the challenge then was how we ascertained what the quality of that was. We worked quite hard with the news media in that space to say, "Actually, here are data sources that you can trust," versus some of the—as you say—enthusiastic amateurs, who perhaps did some things with the data that we wouldn't do or wouldn't necessarily stand behind. So we worked quite hard with the media in that case to understand what was good analysis and what was not, in our view, robust analysis.

Aaron Bell: We found that some of the most informative public communication was by people basically on Twitter, who were doing it alongside their day job, perhaps as a maths professor or something like that. Thank you, Mr Sowden; that is incredibly helpful.

Chair: Thank you very much. We will now go to Rebecca Long Bailey and then Carol Monaghan.

Q200 **Rebecca Long Bailey:** Thank you, Mr Sowden, for speaking to us today. You mentioned that the Government does not have an opt-out for publicly held data but that you had strict rules in place. What challenges has New Zealand faced in ensuring the interoperability of data and data systems, especially across Government Departments?

Mark Sowden: Most of the challenges we have had are around the data quality, the varying levels of quality, and the varying ways in which certain definitions are used. Before we put out a standard on names, we had about seven or eight different ways of collecting names across the public sector agencies. So you ended up not necessarily knowing whether a Mark Sowden on this database was the same as the Mark Sowden in another database. Those data quality issues and data definitional issues have been one of the big issues. What are we doing around that? I am using my Chief Data Steward powers to develop a series of data standards that are compulsory for organisations to adopt—they have some discretion about the timeframe, but it is compulsory for them to adopt them—so that we get all that data at the same quality standard, because what we found, if I

go back to the integrated data infrastructure, was that some datasets simply weren't suitable to be put into the IDI. There were others where we had to do a whole lot of work cleaning—for want of a better term—the data before we could even put it in. So now, what we are doing is going back to the host agencies and saying, "Actually, it's your responsibility to make sure this data meets the right quality standards, and here those standards are."

Q201 **Rebecca Long Bailey:** That is very helpful; thanks. What role have Statistics New Zealand and New Zealand's integrated data infrastructure played in overcoming these challenges and particularly in checking that those agencies are adopting those standards?

Mark Sowden: It kind of goes back to what I said before. We make the decision about what goes into the IDI and we won't put into the IDI anything that does not meet those minimum standards. We will work with those agencies to clean the data up or we may do that ourselves, but we won't put in anything that doesn't meet those quality standards. In terms of, more generally, agencies meeting the data standards, we have not done anything to date, but I am about to pick up the power to conduct individual assurance activity in individual agencies. We have to go into agencies and say, "Are you applying those standards? If not, why not? What is your plan to pick up those standards in time?"

Q202 **Rebecca Long Bailey:** I understand that Statistics New Zealand uses the "five safes" framework to manage safe access to personal data. What insights have you gained from using that approach? Have you tailored it for New Zealand?

Mark Sowden: The insights: it gives the public a lot of confidence. We rely heavily on being able to talk about that, so we can be transparent about what criteria we apply. That gives the public a lot of confidence. We did some research about two years ago, which showed that the public had a high degree of confidence in those five safes. We tested and said, "These are the things we test for." We tested that with the public. What was the second part of your question?

Q203 **Rebecca Long Bailey:** It was about how you tailored it for New Zealand.

Mark Sowden: Partly, we administer it in the context of New Zealand. So is it the right research question? Is it a "public good" research question? We have also tailored it to take on a Māori world view for our Māori population. All requests go through this test as well. It is called Ngā Tikanga Paihere. That is basically a way of taking a Māori world view into those five safes.

At the risk of grossly over-simplifying, data about an individual probably means more from a spiritual and cultural sense to a Māori person than it does a non-Māori person. They put a higher weight on their data about their description; it goes to their real sense of being. So they have a higher degree of protection over that. Using things like the Ngā Tikanga Paihere framework, we are able to say in a New Zealand context—and

particularly in a Māori context—that if the dataset is heavily weighted towards Māori, how do we think about keeping this safe for them?

Q204 **Rebecca Long Bailey:** How easy is it for trusted researchers to access research data? For example, can they avoid filling in multiple forms to access the data, after passing rigorous checks? And what are those checks?

Mark Sowden: I probably have a different view from some of the researchers about how easy it is to fill in the forms and pass the test. This is what we do. One of the things is safe people, safe researchers and safe environments—we only ever check those once. So we will clear the background of a particular researcher. We will look at whether they use data ethically and so forth. Once you have done that test once, you don't have to do it again. A researcher who uses the IDI often will have to apply the test only around the public interest of their individual research programme. So we do try to minimise that.

We also help people make applications, particularly if it is the first time. We have relationship managers who will help the submitters make those applications. They will sit you down and take you through how to fill in the form and what good looks like, and we will go back and forth. We very rarely turn anything down in the end, because we have helped them fill in the form in the first place, so we have made sure the application is robust before we get it.

Q205 **Rebecca Long Bailey:** Lastly, what impact have privacy enhancing technologies made in New Zealand? What role are they seen as playing going forward?

Mark Sowden: To be honest, not a lot here so far. There has not been a lot of adoption of some of those things here so far. I would say not a lot at this point.

Rebecca Long Bailey: That is very helpful; thank you.

Q206 **Carol Monaghan:** Thank you for joining us, Mr Sowden. Some of the evidence we have heard is about the lack of legal certainty on data sharing. In the UK there is a slight reluctance for some data owners to share their data. Could you say a little bit about the role that guidance and legislation play in simplifying rights and responsibilities regarding data sharing?

Mark Sowden: We tend to set very high-level principles in our legislation, and then we tend to go for the guidance being something that is far more easily updated than legislation. That tends to be the more prescriptive. Our legislation tends to be quite permissive and just talks about the sorts of things you need to take into account. We need to take into account Māori cultural interests in data. Hence, the Ngā Tikanga Paihere framework. But that is a good example of where the legislation says we need to take into account Māori cultural interests. To be honest, what those are, and how we think about them, changes over time, as societal norms change over time. We look at the specifics of that through guidance

and good practice, so that you can reflect the changing values of society, and the changing values of New Zealanders, in a far easier way.

Q207 **Carol Monaghan:** So you would say that the legislation provides a framework but there is flexibility within it with the guidance that you provide.

Mark Sowden: That is right, yes. We can reissue guidance. We always do it with the public, and we consult on it—we get public input into it—but we can do that quite quickly to reflect changing circumstances. Covid is a good example, because we changed some our guidance around covid to reflect the fact that, actually, we needed to move more quickly to allow researchers access to data, given the fast-moving nature of covid.

Q208 **Carol Monaghan:** Thank you. What role does AI play in analysing the data that you use in New Zealand?

Mark Sowden: Different organisations use it differently. For research purposes, not a lot. How AI is used, particularly among the New Zealand Government but also among the New Zealand private sector, tends to be more analysis of identifiable data, so that you can work out who to market something to or what services to deliver to particular high-need groups. It is quite specific about individuals. Actually, in a research sense, AI is not used very much here, but we do have a thing called the algorithm charter for when you are using AI in algorithms. Again, this is something that we created under the Chief Data Steward role, and it is designed to bring transparency to the development and use of AI and algorithms. That is about trying to be clear about why we have developed the algorithm, how you might use it, what evidence there is of bias, and what testing we do of bias. The underpinning principle is transparency and having an open book where we can—except for security things—so that New Zealanders can see how that technology is used.

Q209 **Carol Monaghan:** Just to be clear, you have produced additional guidance specifically on AI and how that will operate.

Mark Sowden: Yes, that is right—particularly on algorithms and decision making using algorithms.

Q210 **Carol Monaghan:** Would you be able to share that with the Committee as well? That might be helpful for us to see.

Mark Sowden: Yes, of course.

Carol Monaghan: Thanks very much, Mr Sowden.

Q211 **Chair:** Finally from me—it links to this question of AI—we will be hearing later this morning from Professor Goldacre, who has expressed concerns about the pseudonymisation of data and how that might be susceptible to being cracked and people discovering the identities. AI obviously provides a means to do that. Has this been a concern? You mentioned that the data is anonymised, and I assume you mean that the names and addresses are removed. Perhaps you could just talk us through how it is done there. Do you have any concerns as to whether that is going to be

inadequate for the technologies that can undo that?

Mark Sowden: I think the answer is yes. We have not had any cases of that in New Zealand. If we knew that researchers were going to use some of the AI technology that could take you down that path, it would be one of the things that we would consider when we worked out whether to grant access for that research project in the first place. But yes, you are right. One of our concerns is that, with the increasing use of technology, there is an increased risk of what we call re-identification—being able to find someone in the data. We are addressing that mainly by having penalties. Both in our current law and in the new Data and Statistics Bill, we have penalties for wrongful re-identification and misuse of that data. I don't think you will ever be able to prevent it 100%. We can do our best, but you cannot prevent it. What you can do is catch it. If you catch it and catch someone doing that, there are quite severe penalties. Again, I could not tell you right off the top of my head what they are, but there are quite severe penalties in the law for doing that re-identification.

Q212 **Chair:** Thank you; that is very helpful to know. Just to be clear, it would be a criminal offence to knowingly identify people.

Mark Sowden: Yes. It is punishable by fines, not any prison time, but it is a criminal offence. That is right.

Chair: I see. That is very helpful to know. Thank you very much indeed for your evidence today. It is very good to be able to draw on the experience that you have had in New Zealand. I would be grateful if you would follow up in the way that you have very generously said you would on some of the facts and figures. Thank you very much, Mr Sowden, for your evidence this morning.

Mark Sowden: My pleasure. Thank you.

Examination of witness

Witness: Christian Reimsbach-Kounatze.

Q213 **Chair:** I am pleased to introduce our next witness, Christian Reimsbach-Kounatze from the OECD, who I assume is joining us from Paris. Mr Reimsbach-Kounatze is leading work that the OECD is conducting on the role of data for growth and wellbeing, obviously across a range of countries around the world.

Mr Reimsbach-Kounatze, having been present for the previous session, you have heard some of the things that are on our minds here in the UK. Could you give us a summary of international practice in using data to make particular medical advances combined with the privacy and data security questions that inevitably arise? What do your global inquiries tell you?

Christian Reimsbach-Kounatze: Thank you for giving me the opportunity to present our findings. My name is Christian Reimsbach and I

am co-ordinating the work of the OECD on data governance. I have actually done that since 2012. For those of you who do not know the OECD, it is essentially an intergovernmental organisation that includes 38 member countries—one being the United Kingdom. We look at many different issues around data governance. The organisation itself covers a wide range of policy domains.

One important point to highlight here, because it sets the context for my intervention, is that while I am co-ordinating the work of the OECD on data governance issues, there are many different committees within the OECD that are working on sector-specific issues, including work by the Health Committee within the specific health context. Those are in collaboration with the committee I serve, which is the Digital Economy Policy Committee.

To respond to your question, essentially what we see is an emerging trend towards national data strategies. That is pretty much in line with what we see happening in the United Kingdom with the UK's national data strategy, which is complemented by the more sector-specific health data strategy. We see that also emerging in some other countries.

We also see that the reason countries are moving to those national strategies is that there is a clear understanding that data governance issues need to be addressed by a holistic, whole-of-Government approach. Very often, the concerns that you want to address are not only tied to a specific policy aspect; for instance, clearly when it comes to health data an important issue is privacy and data protection, so you need to involve your data protection authority in policy making and in the supervision process. There is an increasing requirement to take a strategic and whole-of-Government approach in policy making, as that is a clear trend we see happening.

Q214 **Chair:** Thank you very much indeed. Obviously, as you say, countries around the world are grappling with these same problems. Are there any countries that you have studied or that have come to your attention that you think are either in advance of others or doing things in a particularly novel or compelling way?

Christian Reimsbach-Kounatze: First, I should highlight that we have actually started the work in the past, and we did a country review four to five years ago assessing what countries had been doing. We are now in the process of redoing that survey. I say that to highlight that our view of what countries are doing is not complete. Nevertheless, we have begun to see what some of the frontrunners are doing, and what innovations are happening. One thing that we noticed is that the UK is one of the leaders or innovators when it comes to what is happening, not only because of its national data strategy that the UK has implemented, among very few countries, but when it comes to some specific aspects.

To reply to your question, it is hard to pick just one country and say, "This is the country that has done everything very well." There are countries that are really doing well when it comes to specific aspects, and I am sure

we can go to those specific aspects, while others are probably struggling on those aspects. That is why in our work we are trying to identify the different policy issues that come with data governance and to identify the best parties for each of those policy issues. It might be more helpful in our discussion today if we could focus on those specific policy issues and then highlight what good practices could be.

Q215 **Chair:** Your perspective on this is very valuable, given that you look around the world—obviously, OECD members—on this. Before I turn to Aaron Bell, when we might ask who the leaders are in specific aspects of this, would you say that there are any global frontrunners generally?

Christian Reimsbach-Kounatze: It is difficult to say definitely that there are global frontrunners—

Chair: That is when it gets down to specifics—that is understood.

Christian Reimsbach-Kounatze: But what we can say is that there are certain countries that are definitely being looked at, and the UK is one of them. The other country that has often been a point of reference, let us say, is Australia. When it comes to private sector data and what the Government have been doing to facilitate data sharing within the private sector, one of the countries that has also been referenced a lot is Japan. When it comes to health-related data more specifically, Korea is another country that has also been referenced a lot, and the work of the European Commission is also being looked at, not only in the EU member countries, but globally, because it has implications outside the EU.

In certain respects—again, it depends on the policy domain that we are looking at—you might also look at the United States and Canada for certain aspects. I will just highlight the fact that Canada is also in the process of developing the national data strategy specifically for the health sector. We also see interesting developments happening there, but again—the devil lies in the detail—it really depends on the specific policy issues that such initiatives are trying to address in order to say whether they can really be identified as "best practice".

Chair: Thank you. That is very clear.

Q216 **Aaron Bell:** Thank you, Mr Reimsbach-Kounatze, for your time. I will go into some of the specifics, starting with opt-outs. In the United Kingdom we have a national opt-out, by which people may choose to opt out of having their data used at all. We have just heard that in New Zealand they do not have that, but they have a right to opt out of it being shared with third parties. How typical are such opt-outs around the world? Which other countries, if any, have them?

Christian Reimsbach-Kounatze: Thank you for the question. First, on the specific question on how countries are doing in respect of opt-out, this is something that we are in the process of assessing, in particular as I mentioned with the Health Committee. I am not able to tell you exactly how many countries have specific opt-out regulations.

This is something that has been emphasised in a number of our council recommendations, but for those of you who do not know about the OECD council recommendations, essentially they are one of our major legal instruments that the OECD produces. They are not legally binding. Nevertheless, Governments take those legal instruments very seriously, because they represent a political commitment to implement the principles quoted in them.

In all our major council recommendations—here I highlight the recommendations on health data governance and on enhancing access to and sharing of data—there is no specific recommendation on whether to go primarily opt out or opt in. However, what is clearly articulated in the recommendations is that individuals—data subjects—should have the possibility of withdrawing their consent once they have given their consent. That is the standard at the moment when it comes to that particular question. Whether Governments have an opt-in or opt-out preference as a default position, it is hard to say what is better. It depends on different circumstances.

Q217 **Aaron Bell:** Most countries have some method of opt-in. We have just heard that in New Zealand, from a Government perspective, they do not. The data is available for Government use for everybody in New Zealand, which is established in an Act of Parliament. Is that unusual? Would most countries approach this issue with some measure of opting in or opting out?

Christian Reimsbach-Kounatze: Based on the information we have, I am not in a position to assess whether or not it is unusual. As I said, we haven't completed the work, in particular when it comes to health-related data, which I think is what you are interested in. That is undertaken by our colleagues in the Health Committee, so I would have to ask them if they have a specific view on this kind of question. I am happy to come back to the Committee with responses on that specific question.

Q218 **Aaron Bell:** That would be very helpful. It sounds like there is some work ongoing there as well. Obviously, we give people a legal right to opt out, but the more people opt out, the less useful the data is. Do you have any insight as to how other countries manage opt-outs to try to persuade people not to opt out, and persuade them to opt in?

Christian Reimsbach-Kounatze: The key point is the issue of trust. I am pretty confident that we may come back to that point. It all depends on the level of trust that citizens or, let us say, data subjects have in the institutions that are involved in the whole data ecosystem. That depends on a number of different factors. Giving individuals control is definitely an important element of that trust. You could make the argument that, yes, if we give people the control to withdraw their consent or to prohibit or prevent the use of their personal data, it may actually present a risk that the data may be underutilised. On the other hand, it may be a condition for enabling trust in your society in respect of how data is used.

From a policy maker's perspective, the question is: what is the key lever that you want to use? From the OECD perspective, and this has been articulated very clearly in our council recommendation on enhancing access and sharing of data, an important element is to ensure trust and that the data ecosystem is considered as trustworthy. Transparency is one important condition, but also putting individuals in control over the data is another important one. Consent and the ability to opt out or in are important parts of the toolbox in that whole discussion.

Q219 **Aaron Bell:** Thank you. You just mentioned transparency, and that is what I want to ask about next. We have already heard in this inquiry that there is a need for transparency in how data is collected, who it is shared with and what it is going to be used for. What initiatives or approaches would the OECD recommend for transparency? Are there any particular examples of best practice from other countries that we could learn from?

Christian Reimsbach-Kounatze: As I highlighted, I pointed to the council recommendation on enhancing access and sharing of data. If I may, I will quickly quote the specific provision, because I think the wording is very important. Essentially, the section on reinforcing trust across the data ecosystem "recommends that Members and non-Members having adhered to this Recommendation… empower and pro-actively engage all relevant stakeholders alongside broader efforts to increase the trustworthiness of the data ecosystem". Essentially, adherence should therefore "Enhance transparency of data access and sharing arrangements to encourage the adoption of responsible data governance practices throughout the data value cycle".

I am quoting this because there are a number of concepts here that are important. An important one, because it is actually often overlooked, is the need to increase transparency over the entire data value cycle. When we talk about the data value cycle, we mean the whole process from data collection to data processing and data storage, and also eventually data deletion, which may in certain cases actually be necessary.

How do you achieve that? What Governments are doing is essentially to inform the public, and there are many mechanisms for doing that. The key challenge that we see is that obviously you are eventually dealing with a large population, and informing a large population can introduce huge transaction costs. Certain countries use representative groups, depending on the context, and others only really involve the individuals concerned. This all obviously depends on the scale of the data that is being used for whatever initiative or effort you are talking about.

Q220 **Aaron Bell:** Sorry to interrupt. Are there any particular countries that you can point to that are getting this right?

Christian Reimsbach-Kounatze: One country we have looked into in our efforts is Australia. We heard from an Australian representative that one of their key lessons learned, particularly given their history as a key mining country, is always to involve the public when it comes to large-scale initiatives such as mining. They consulted the public and put in place

sophisticated consultation mechanisms. What is interesting is that they taught us that they have transposed those mechanisms in the context of data to exploit and govern the use of data as a national resource, so to speak. That is an interesting example that I would like to point you to.

Q221 **Aaron Bell:** We have talked about giving people control and about transparency. The third pillar is engagement on data sharing. What types of engagement activity are most effective for gaining public trust? Again, are there any international examples that you would like to draw our attention to?

Christian Reimsbach-Kounatze: It depends on the kind of stakeholder you are aiming to engage. We have talked a lot about engaging citizens or data subjects, who are obviously key actors when it comes to trust in the data ecosystem. Private sector actors also need to be engaged, and we see that happening not only in the UK but in other countries. Depending on the type of initiative you are talking about, that may be a crucial element. There are also some key criteria. Again, I point you to the council recommendation on enhancing access, which encouraged data-sharing partnerships. I note from the UK example that that is an important element. The key criterion that is emphasised in the council recommendation is to ensure that data-sharing partnerships are competition-neutral, meaning that they do not distort competition by favouring large incumbents or national champions. That is an important element that we have been looking at.

In terms of best practice in this area, most countries have specific rules on how to address data-sharing partnerships. We see that happening in France, for instance, where they have initiatives in that respect. However, so far we have not been able to assess the extent to which those countries are really making sure that those initiatives do not disturb competition. We will be digging deeper into that as we do our assessment.

Q222 **Aaron Bell:** Finally, a lot of witnesses have raised the experience of the covid pandemic and how it relates to digital data, in terms both of tracking the spread of the disease and of the trials that have been done on vaccines and therapeutics over the past couple of years. Are there any countries that you would draw our attention to that have got covid right from a digital data perspective?

Christian Reimsbach-Kounatze: Sure. We have indeed done an assessment of what countries are doing in their reaction to covid. One general trend that we observed was that countries in Asia, because of their experience of other pandemics or infectious diseases, have been in a better position to take advantage of the pre-existing regulatory environments. One of the countries very often mentioned is Korea, in relation to its ability to deploy apps, making sure that data is used and linked between different institutions, and also the speed at which data is exchanged between the relevant institutions. I am happy to send you details on the Korean example, because there are many different layers to that—I have the information right in front of me, but it would take too much time.

Q223 **Aaron Bell:** That would be helpful. Do you think that public trust was built by the handling of previous pandemics and that that was a carry-over benefit to those Asian countries?

Christian Reimsbach-Kounatze: Yes, but at the same time I think it is also important to acknowledge—we may get to this point—that there is also a cultural dimension to that trust. Depending on the country, we know that trust in public institutions can vary from one country to the next. This is also an important element to consider. I would say that the Korean population, as you rightly summarised, had a positive experience of the way their Government handled previous pandemics, and therefore they were able to embrace the recent initiatives and implement this as fast as they were able to do it.

Aaron Bell: Thank you.

Q224 **Rebecca Long Bailey:** Thank you for speaking to us today. We have been told previously that one of the main barriers to data sharing in the UK is the lack of interoperability of datasets and systems, especially across Government Departments. How big a problem is this for other OECD countries and what is its impact?

Christian Reimsbach-Kounatze: Data interoperability is an important challenge. One thing that we rarely talk about is the different dimensions of data interoperability. I say this because it is important to break down the problem in order to understand what different countries are doing and how they are addressing this. Very often—I think this is where most countries are focusing when it comes to data interoperability—it is more focusing on the technical aspect and, in particular, on what is often called synthetic interoperability, which means making sure that we use the same data formats.

This is an old problem that the Government have been trying to address for a long time. Significant progress has been made because there exist sometimes sector-specific data formats and so on that we see being promoted across different policy domains. Essentially, I would say that almost all the policy initiatives that we have looked at that deal seriously in trying to address the issue have actually articulated provisions that address the issue of data formats.

Where there are shortcomings in other dimensions of interoperability, one is often referred to as semantic interoperability, which means we make sure that not only the data formats can be read by different systems, but that we also attach the same meaning to the different formats. Currently, we see that Governments—including, from what I have noted in the UK, where there are challenges—are making efforts to address the challenges. Very often the problem is due to the fact that semantic standards are not really established. If they are established, as they may be in a certain area, there might be even a proliferation of too many standards, and therefore it becomes an issue of how we decide which standard to use and agree on, so that is one important dimension.

In some countries—the UK being one but also Australia, and we heard about New Zealand—what we note, and this is also happening in a number of EU countries, is that there is a trend towards what is called trusted custodians of the data: essentially, a trusted actor within the ecosystem who is responsible for co-ordinating not only the policies but the data sharing mechanisms. This actor typically performs the role of standardising or converging towards standards for data interoperability, including semantic interoperability.

I will conclude with a third point, on which, I would say, a lot needs to be done: the area we refer to as legal interoperability. That is particularly important when you ask, "How do we ensure that data is also exchanged across borders?", because one thing we learned from covid is the importance of ensuring that data that is produced in one jurisdiction can also be used to inform policy making, regulations and measures in other countries. However, for that to be possible—obviously privacy is perceived by many actors as a barrier, but another important point attached to that is on ensuring that the rights and obligations that are attached to the data are actually really attached to it when the data is being moved around. That is what we refer to as legal interoperability—or policy interoperability, you could say—and that is where, at least for now, very few countries have made efforts. I would not actually be able to point you to one specific example, at least not yet—hopefully, but not yet.

Q225 **Rebecca Long Bailey:** Thank you. Moving on to the issue of privacy, to what extent can a legislative and regulatory balance be struck between enabling data sharing and protecting an individual's privacy? Have any OECD countries come close to achieving that balance, and why have they achieved it?

Christian Reimsbach-Kounatze: Maybe to answer your question with a straight answer, I think there are a number of countries. There is a clear trend in countries being able to continuously better achieve that balance. The reason is because our countries, in particular, now have at their disposal a wide range of tools to basically strike a balance. We talked in previous interventions about the role of privacy enhancing technologies, or trusted data environments. Those are basically technical/organisational approaches to strike that balance, because essentially you cannot use the data while still protecting the rights and interests of individuals.

Another important aspect to that is what I referred to as this kind of holistic approach to data governance, where you do not just focus on preventing and avoiding the risks, but are actually also focusing on, "How can we make use of the data for the public good?" I think that kind of shift in policy attention and objective has been a key enabler for our whole discussion.

Of all the countries, essentially, that are moving towards national data strategies—besides the UK, obviously, because I am sure you are interested in the other countries—I would point to Norway. It was the first country that started to think about a national data strategy. Norway was interesting because, before countries started talking about a national data

strategy, it actually had ambitions to implement something like a national privacy strategy, which I would perceive as a kind of predecessor of what we now refer to as national data strategies. I hope that provides the answer to your question.

Q226 **Rebecca Long Bailey:** Thank you; that was really helpful. I have one final question. The UK Government announced yesterday that they would lay a Bill before Parliament to cover issues such as lessening the burden of data protection on businesses and institutions, and also reforming the sphere of data sharing. Obviously, we have not seen the Bill yet, and it will cover a lot of issues, but from your experience of the UK environment, what key things would you like to see within that piece of legislation to ensure that we do get that balance between being able to effectively share data and protecting people's privacy?

Christian Reimsbach-Kounatze: Before I answer the question, it is important to highlight that at the OECD we rarely comment on pending legislation, except when we are invited by Governments to consult. Also, having not read it, I am unable to answer the question. However, what I can say—although it may be obvious—is that there is the expectation that whatever the legislation, and however it is articulated, it needs to be in line with the council recommendations of the OECD, including the recommendation on enhancing access to and sharing of data, the recommendation on health data governance, and the OECD privacy guidelines, which have been a reference for privacy around the world. Those recommendations are actually setting the boundaries of what good policies would look like.

Rebecca Long Bailey: Thank you. That is very helpful.

Q227 **Carol Monaghan:** We have had a lot of evidence from witnesses who are pretty much unanimous in telling us that a trusted research environment is central to safe and secure data sharing. What should the UK be learning from other OECD countries about building and supporting a trusted research environment?

Christian Reimsbach-Kounatze: There are different dimensions to it. One is that it is important to understand that those trusted research environments fall within a range of what we call privacy enhancing technologies and methods.

In our work that we are currently doing on privacy enhancing technologies, we have identified what the UK refers to as trusted research environments as an important element of what we classify as data accountability tools. That is a broad family of tools that includes personal data stores and those kinds of environments, but at the same time there are other tools. From what we see, there is a need to think through all those methods within the broader context that there are those different tools. I am not aware of how exactly the trusted research environment operates in the UK, but there is obviously the expectation that those environments will use the full tools available—technical tools as well as

organisational tools—to take advantage of data processing, gaining the benefits of data while protecting the interests of the individuals.

In terms of what is best practice or where we can take lessons learned from, interestingly, one of the institutions or groups of actors that has been leading in that field is the national statistics agencies. Therefore, to me, it is not a surprise that we also had a witness from New Zealand from that community. Around the world, and obviously in the OECD member countries, national statistics agencies have developed the competencies to deal with those environments, because they had control over very sensitive microdata, and those data needed to be used for research purposes and the like.

It is therefore also no surprise to us, at least from that perspective, that you have countries such as Australia that have basically assigned their national statistics agencies as one of the key trusted agents in the whole policy on data sharing; it is exactly because of that knowledge. I am not fully aware of how the UK is organising it and what the role of the UK's national statistics agency is in the landscape—this is something we are currently looking into—but I would assume that it would probably play an important role as well.

Q228 **Carol Monaghan:** We have also been hearing about privacy enhancing technologies. To what extent can data be properly anonymised before it loses its usefulness—for example, in terms of its linkages to other data?

Christian Reimsbach-Kounatze: This is obviously an important question. I think in the previous discussion, the question was also raised about the role of AI in undermining some of the efforts to anonymise or pseudonymise datasets, and it is definitely an important aspect to consider in the discussion.

A lot of progress has been made in the field of privacy enhancing technologies. In our work, we are looking at—as I already said—different classes of privacy enhancing technology, one of them being what we refer to as data obfuscation tools. That includes approaches such as federated learning, where you basically send the algorithm to the data, or differential privacy, where you introduce selected noise elements into the data to mask elements of the details of the individuals. All those tools, and the sophisticated manner with which actors are learning to use them, are enabling institutions to use the data while still protecting privacy in a way that was not possible in the past. Progress is being made continuously in this field.

Q229 **Carol Monaghan:** Thank you. Maybe I could move on to AI. AI obviously has great potential in terms of analysing huge amounts of data. There are also concerns that it might lead to biases or arbitrary decision making. How can we ensure that AI analysis of data is transparent, unbiased and open to scrutiny?

Christian Reimsbach-Kounatze: This is indeed an important question. I would like to again point to the council recommendation on enhancing access and sharing of data, which addresses the issue from two

perspectives. There is, let's say, the social dimension, and I have already pointed to the importance of social participation. This issue has been articulated by one of our member countries, Canada, which was basically concerned about the fact that minority, in particular indigenous, populations were not well reflected and that there were biases in the data towards those social groups. The question is, how do you address that concern? One of the solutions that is articulated in the council recommendation on enhancing access and sharing of data is to make sure that there is a sufficient level of participation from all the relevant social groups, including minority groups, which is the area where there are probably problems for many different reasons. It is important to address that.

The other issue, which we talked about previously, is transparency. Transparency is important on a data level when it comes to AI, because citizens need to know what kinds of data on them is being used. Transparency is also important on an algorithm level. This is where we see some potential problems and tensions. Very often those algorithms may be proprietary by nature, because they are owned and protected by intellectual property rights. Therefore, revealing the inner details of those algorithms may violate or conflict with commercial interest. The question is how you achieve that. There are many different approaches. It also depends on whether the Government have been funding those algorithms.

Q230 **Carol Monaghan:** Are you able to point us to particularly good examples of this?

Christian Reimsbach-Kounatze: I am happy to point you to some examples. One thing I can already say is that, in the area of science and research, where Governments and the public have been the main funders of the undertaking, it is much easier for the Government to push for opening access—not only to the data but also to the algorithms. The situation becomes much more complicated when there are private-public partnerships involved. I am happy to point you to some of the initiatives that we are essentially digging out, and I am happy to send it to you. The issue here is much more complicated, as you can imagine, because of the entanglement of interests when it comes to those partnerships.

Q231 **Carol Monaghan:** In the UK, the Office for National Statistics has developed the "five safes" approach to promoting best practice and data sharing. What is the OECD's approach to ensuring that personal data is shared securely and ethically?

Christian Reimsbach-Kounatze: I would say that the "five safes" nicely summarise the key elements of all our recommendations. In other words, all the elements of the "five safes"—safe people, projects, settings, outputs and data—are articulated in our recommendations, because they make sense. It obviously makes sense to ensure data quality. It absolutely makes sense to ensure that those accessing the data are trustworthy, that the use of the data is for the benefit of the public, and that it is ethical.

All that has been articulated, although not necessarily by explicitly referring to the "five safes". I would argue that there are additional issues that need to be taken into account, but those "five safes" are definitely articulated in our recommendations—through other words, but they are included.

Carol Monaghan: That is really helpful. Thank you very much.

Chair: Thank you very much indeed for giving evidence today. Your perspective on a wide range of countries, which is clearly ongoing, is very important to our inquiry. We would be grateful if you followed up on some of the particular aspects you mentioned, but we are very grateful for your time and consideration today.

Examination of witness

Witness: Professor Ben Goldacre.

Q232 **Chair:** I am pleased to welcome, appearing in person, Professor Ben Goldacre. Professor Goldacre is the Bennett professor of evidence-based medicine and director of the Bennett Institute for Applied Data Science at the University of Oxford. He was commissioned by the Government to undertake a review of questions of data sharing—the Goldacre review. Its report, "Better, broader, safer: using health data for research and analysis", was published just last month, on 7 April.

Thank you very much indeed for coming, Professor Goldacre. We have lots of questions, but perhaps we will start with a question of context. Last summer, I think it was, there was a programme to allow people to opt out of sharing their GP data—it was GP data for a planning and research programme. A surge of people wanting to opt out—I think over a million in a short number of weeks—caused that programme to be paused and perhaps even abandoned. That was obviously in your mind when you were conducting your review. Would you perhaps reflect on that episode and share with the Committee your thoughts on what went wrong?

Professor Goldacre: I think that cuts to the core of all the challenges and opportunities around working with this kind of data. The GP dataset is probably the single most valuable NHS data asset that we have, and that is because of two things. First, there is its granularity: it has a lot of detail about each individual patient. There is something in there about almost every health service contact and every prescription, blood test, referral and diagnosis that is recorded in primary care, but there is also information about what has happened to patients in secondary care, albeit a little more sparse.

The data is also very complete in its population coverage—it covers the whole of the country. That really makes it the jewel in the crown of NHS data. It is one of the things that makes the NHS, and the UK, such a unique place to do this kind of work. It has huge opportunities for

traditional academic research; for analysis for service improvement to monitor the quality, safety and effectiveness of care; and for driving life sciences innovation.

However, because of the granularity and the comprehensiveness of the coverage, it also presents substantial privacy risks to patients. Like all NHS datasets, it contains information that many patients—although not all—would prefer not to be widely accessible to people outside of trusted clinicians or family members. As the dataset is so complete, it poses additional privacy risks. The challenges and the shortcomings of approaches such as pseudonymisation—removing names and addresses from datasets—are exacerbated the more comprehensive and granular a dataset becomes.

In some respects, you could look at data as being a little like nuclear material. In the past, people have said that data is like oil—the oil of the 21st century—which implies that it is just sitting there waiting to be used or burned. In reality, it is a lot more like nuclear material in the sense that when you first access small bits of it, it is not actually very useful; it needs to be refined and processed. After it has been refined and processed, two things happen: first, it becomes tremendously powerful, and secondly, it also becomes rather dangerous. Once it has leaked, it cannot be unleaked, and in order to do good with it, you have to work very carefully with it, while minimising harms.

The current paradigm—the way that we work with these large datasets—is to take names and addresses off in a process called pseudonymisation, and to distribute the data to multiple different locations. Taking names and addresses off, as I hope we will discuss in a little more detail, presents quite serious shortcomings when it comes to protecting patients' privacy. I think a number of patients and professional groups were concerned that sufficient privacy mitigations had not been put in place before the programme was initiated. In that sense, it was, disappointingly, a rerun of exactly what we saw in care.data in 2013, which was exactly the same project and failed on exactly the same terms, with around 1.5 million opt-outs at that time as well.

I am very pleased that, after we first started raising concerns about the shortcomings of pseudonymisation and the power of working in trusted research environments to mitigate that risk, a letter from Ministers to GPs in the middle of last year made a promise about the restart of the GP data programme, which was that it would continue only after they had built a national trusted research environment that could hold the data securely and make it accessible to all legitimate users, while mitigating the risks. I am confident that by doing that, you can not only mitigate risks, but begin to earn public trust.

Q233 **Chair:** Thank you. As you said, we will go into more detail on pseudonymisation—my colleague Zarah Sultana will ask some questions about it. Do you think the public and the trusted experts who you mentioned were right to be concerned about the programme that was in operation?

Professor Goldacre: Pseudonymisation and dissemination have such serious shortcomings that they present privacy risks, but I do not think we are necessarily sitting on top of an enormous backlog of privacy disasters. The problem really is that, because we haven't addressed those privacy risks by building trusted research environments, and because we have taken a rather chaotic and ad hoc approach to access to data through dissemination, we now see a lack of people doing good work at all in data.

Q234 **Chair:** On this point about those concerns, it was communicated widely through the media and social media, and a lot of people were signing up to withdraw their consent. Do you think they were right to do so?

Professor Goldacre: I think it is a mistake to launch an enormous programme like that without first of all making it clear to patients exactly what the mitigations will be. I think the national GP dataset is so granular and so comprehensive in its coverage that it wouldn't be appropriate to share it outside of the TRE. So overall, yes, I would say it was wrong to try to launch a programme like that before we had made a commitment that all data access would only be through a trusted research environment. But that, as I understand it, is now policy, so we have a very positive future ahead of us, in part because trusted research environments don't just mitigate privacy risks; they also make for a much more productive environment for working with data.

Q235 **Chair:** Indeed, and we will come to that. Did you withdraw your consent for your data to be used?

Professor Goldacre: Oh, golly. I did, because I know so much about how this data is used and how people can be de-anonymised, and also in part because in the past, to a greater extent, I have been in the public eye from doing public engagement work and I have friends who have had their data illegally accessed through national datasets, not health datasets. I suppose because I work in this field, the risks are very salient to me.

Similarly, in my personal experience, friends who work in emergency planning are more likely to stockpile food, for example during the Olympics in 2012. So, to an extent, risks that you work with closely have a greater emotional salience. But yes, I did. I am only hesitant because I have not said that before.

Q236 **Chair:** No, no, clearly—that was my next question, really, given the importance of this. You will have seen from previous evidence sessions that this Committee does not doubt the huge importance of being able to analyse data and to make the kinds of breakthroughs that we do. That makes it especially important that there is a regime, a context, in which people can have confidence. Clearly, that episode shows that at least a million people and more, including someone as expert as you, did not have that confidence and therefore it was a setback. How was it, given all the fine minds in NHS England and the Department of Health and Social Care, that such a blunder was made?

Professor Goldacre: I cannot speak to the decision-making process that effectively led to the relaunch of care.data. I can say, from my experience

of engaging with Government, albeit from an adjacent role, such as the one I have, that what I saw up close, being candid, was a kind of groupthink that tended to overstate the privacy benefits of pseudonymisation.

I think that was partly driven by a long-standing lack of technical expertise among people in senior roles, working on issues such as data architecture and analysis, both for research and for planning. I think it was also driven in part by the fact that, in the legislative regulatory frameworks that we use to gain access to data, in order to avoid them outright obstructing good work with data, there is a degree of grey area that has to be used in our interpretation of words like "anonymised" and "pseudonymised". Reading between the lines, there has been a bit of a sense that we shouldn't rock the boat too much and problematise our interpretation of those terms in some of the legislation and regulations.

I think it is reasonable for us to have access to patient data through secure mechanisms, and I actually think that data dissemination is acceptable in some circumstances—for example, where the patient has given consent for that data to be sent elsewhere or where there are very extreme alternative mitigations, such as data minimisation sampling. Then there may be cases where it is safe to share the data via dissemination.

But overall, I think there was a kind of groupthink. I am very pleased and reassured at how quickly that was eradicated after we started raising concerns—and others raised concerns in the system, too. I should say that I am aware of others in similar senior roles who have been anxious about their own data being accessed. When I and others involved in the review started raising concerns about this, it was striking that often people in more junior and technical roles in the system put their hand up or emailed or called to say, "I am actually really glad that you have said that, because I have been surprised at the extent to which people haven't really engaged with the shortcomings of these approaches before."

I hope those misunderstandings are a thing of the past. I feel very confident that by adopting the TREs as policy, we are moving forward to sunlit uplands of not just more secure working methods but more efficient working methods.

Q237 **Chair:** Thank you. That is a very valuable and fascinating insight. You may know from the work that this Committee and the Health and Social Care Committee have done on covid that we identified a problem of groupthink where people with different views are unable to challenge what seemed to be a consensus. That seems to have been the case here but, as you say, through your review and what the Government have said about implementing it, there is a chance for a reset.

Professor Goldacre: I do think additionally that there is a further challenge. Throughout the review, we talk about how there is no real sense of a centre of gravity or status around people who build platforms, people who do data curation and secure analytics. That has been resourced, in academia in particular, via funding people doing single

clinical academic questions in single academic papers. Because of that, we have not had a commons of knowledge or a body of expertise thinking about these kinds of privacy risks and their mitigations.

One key recommendation in the review is that UKRI and NIHR should engage robustly with strategic, but also open and competitive, funding for people who have expertise in privacy, in security risks and their mitigations, and also efficient methods of curation and secure analytics in data. It is the neglect of that intermediate work that has really held us back, both for privacy management and for efficient analytics with health data.

Q238 **Chair:** Thank you. Finally, before I turn to Zarah, I want to reflect on GP data incident last summer. One thing that caused it to come to a head was that there was initially a very short period for people to opt out. Was that, in your view, an attempt on the part of the authorities to bundle this through before people noticed, and to have people signed up by default? Do you think that was deliberate and part of the intention?

Professor Goldacre: I couldn't possibly know the motivations for it. It is possible that, in part, it was driven by a technical challenge. If you are granting data access by gathering data and then disseminating it to multiple different locations, that is in large part an irreversible act. You cannot allow people to reverse their decision to opt out where you are disseminating data out to multiple locations; it becomes practically and technically extremely hard to manage.

It is probable that that drove the sense of, "It's right now or never that you have to opt out, and it is going to go ahead." As I understand it, the new thinking around opt-outs is that it is important that people will be able to reverse their decision in future, and that that will be upheld, but also, of course, that data will be accessible only in trusted research environments. That will in large part mitigate the risks and allow us to earn the trust of the public, professionals and campaigners, so that I would hope people would not feel any need to opt out.

Q239 **Chair:** Thank you. Perhaps I could give you a chance to comment on the more charitable interpretation that, rather than trying to bundle it through on the quiet—

Professor Goldacre: I think that is the true interpretation.

Q240 **Chair:** But given the context of last summer and covid, was there what I would portray as the more charitable interpretation, that the benefits of sharing data and making rapid progress in the context of covid was thought to justify what turned out to be an attempt to cut corners in a way that didn't work? Do you think that was a motivation—to cut through some of the delays to sharing data, to help manage the pandemic? Was that a motivation that you detected through the conversations you must have had with people engaged?

Professor Goldacre: I think there may have been a sense that we should get on—that it is now or never. The bigger problem was that there wasn't

sufficiently widespread understanding of the shortcomings of pseudonymisation and dissemination, so people thought, "This is the only way we can work—through pseudonymisation and dissemination. It will always be controversial, so you might as well press on." That may have been the reasoning, but as long as we move towards secure and efficient working in TREs, and as long as we crack good delivery of those TREs, which is a separate challenge, all these challenges are hopefully a thing of increasingly distant history.

Chair: Indeed; thank you. That takes us very neatly on to Zarah Sultana, who is going to ask some questions about pseudonymisation.

Q241 **Zarah Sultana:** Following up on people's privacy concerns, I noticed at the time that there were concerns about data being shared with third parties, including private companies, and people fearing that their health data was going to be used for profit. Do you have any insight on that being a concern among all the others?

Professor Goldacre: It is a really important issue. My personal view is that it should be okay for commercial organisations to innovate by using NHS data. However, I don't make choices on behalf of the community; that is the job of this organisation—Parliament. I do think that if we use trusted research environments, we can do something really important, which is to separate privacy concerns from the political or moral judgment about whether private companies should be able to profit from NHS data. We can say it is possible for a large organisation to work on your data, to make money from it, and to develop new techniques, tools or devices by using that data. We can allow them to do that without at the same time worrying that they will be able to see our medical records. After we have adopted trusted research environments and competently implemented them across the community, we can grant access to many more people. Then it becomes a pure moral or political decision about who we grant access to.

My personal view—it is only my personal view—is that the NHS should grant access to commercial organisations but that we should take a strong strategic approach to ensuring that we get equity in any innovations that are derived from NHS patient data. But it is only a personal view, and it is not a thing I think about very carefully. The thing I think about most is how to ensure that people can get secure and efficient access to data.

Q242 **Zarah Sultana:** In terms of the scale of the challenge on refining the data, it requires curating, managing, cleaning and preparing this dataset. How long, hypothetically, are we talking about from a starting position? If we move forward with this, how long would it take?

Professor Goldacre: Curation is a separate challenge, and it is an enormous challenge, because NHS data was not made for researchers or analysts. It is essentially an aide-mémoire for clinicians like me, and patients like everyone, to keep track of what has happened, to help make informed decisions about treatments in the future. You have to take that data—it is often described as messy, but I think that is unfair, because it is

made for a different purpose—and process it into the shape of data that you want for your analysis.

If we look at the GP data, there is not a variable that says, "Does this patient have or not have diabetes?" There are multiple single clinical events where a GP has chosen to type, "This patient currently has diabetes." There are multiple different types of diabetes. They might say pre-diabetes, which might mean something different to different people. There will be patients who are on anti-diabetic medication and who have never had a formal diagnosis of diabetes recorded. There will be people with a formal diabetes diagnosis who do not seem to be on any treatment, so they are probably less severe or have a different type of diabetes. You will then look at their medication records, blood tests, referral records and out-patient history. From all that complicated administrative data about individual clinical events, you will try to boil that down to one variable about one patient, saying, "For the purposes of the analysis that I'm doing today, I'm going to say whether this patient has diabetes or not."

The ABPI told us—and I agree, from many years of working with this kind of data—that they think that about 80% of all the work in a data-analysis project working with NHS data is spent on data curation, so they have recommended that 80% of the money that gets spent on making data accessible should be spent on data curation.

Historically, that job has been done in a very duplicative and piecemeal way, in part because of data dissemination. The same national datasets—cuts of GP data, not national data, and the whole of HES, the hospital episode statistics, or any of the other national datasets—get sent out to hundreds, perhaps thousands, of different locations. They get stored in different computational environments, different data schemas, differently named columns, and slightly different cuts of the underlying data.

As well as the duplicated costs that that entails, it means that the work that people do to process that raw data by writing computer code, the work that people do to curate that data, is not portable between different settings. Also, because UKRI and NIHR have not foregrounded data curation and secure platforms as a high-status independent activity of its own, that work is low status. It's often hidden and often thrown away. People think it's not very important or very useful. It's also often actively withheld. I know of groups in academia, for example, that think, "Well, actually, my data curation pipeline is part of what allows me to deliver analyses quickly, but not other, competing academic groups." Sometimes people don't want to share even the simplest code lists that they use to process that data.

That is the starting position. However, first of all, what we do have across the whole community, across the whole ecosystem of people who work with data, is a lot of very talented individuals with a lot of practical knowledge of doing this work, which is currently hidden. In the data curation chapter of the review, there is a lot on how we can capture that prior knowledge first of all. Secondly, if we get everybody working with the same datasets in the same environments, all their data curation work will

be reusable between different groups. You move to a culture where everybody works in the same place. They arrive to find other people's curation work that they can reuse or modify, or review and improve, but also, when they do their own curation work, they leave that behind, so everybody who works with the data leaves the data in a better state for everybody else.

You also need to employ people to do two things. First is to create standard approaches for data curation, because at the moment what happens is that if everybody does it in a different way, it's a bit hard to read and reuse everybody else's work. Also, you need to pay a not very large number of people to sit down and do curation work on some of the high-value or high-use datasets as pioneers, if you like, to learn and prove out the best methods of doing that, and get them working using standard curation methods.

If we do that—it's not a big job; we're talking about dozens of people rather than hundreds—the enormous hidden costs of the duplicated effort that we have at the moment could rapidly be eradicated. Then, I think, it's a matter of a couple of years before you can create an environment where people arrive to find open code that allows them to do their job quickly and that meets a lot of those challenges.

Lastly, it is worth reflecting on the fact that data curation, because it is hard and time consuming, is often done in a hurry and quite badly, and that means that you cannot trust the results. There was a lot of fantastic discussion in this Committee on a different set of challenges, around open and reproducible science. I think all the lessons of that are directly applicable to NHS data curation.

I think that if we can surface prior art, if we can get UKRI and NIHR to engage with this in a serious way and if we can avoid the historical approach of black box analytics, where nobody is allowed to see the work that is done, or where the data curation work is captured with vendor lock-in, using proprietary data curation methods, then we can very quickly have a very efficient and productive landscape.

Q243 **Zarah Sultana:** Thank you. That was really insightful. Your review made 185 recommendations—incredible. It highlighted privacy concerns, as you said, around pseudonymised data, and you recommended that that should be mapped across the health data ecosystem and then closed down, so what techniques do you think could de-identify personal health data without rendering it less useful for research and innovation purposes?

Chair: May I ask you to be succinct, Professor Goldacre? We have lots of questions for you. You are very expert and we are keen to get all the subjects covered.

Professor Goldacre: Okay. First, it is important to understand the risks in pseudonymised data. You take off the names and addresses, the direct identifiers, but you can still identify an individual if you are a malicious

attacker, and we have to think about threat models and risk mitigation. You can identify people because of other things that you know about them.

The classic example that appears in security engineering textbooks is that you could re-identify Tony Blair in health data because you know the approximate dates in which he had an abnormal heart rhythm reset while he was Prime Minister. By knowing the week in which that happened, the kind of procedure he had on two dates, and his approximate age and location, you could probably find only one person with those characteristics. Having found the unique identifier for that person, you can then see everything else in their record. Women are particularly at risk, in my view, because childbirth is something that appears in your medical record and is typically known by colleagues, people at the school gate and so on.

We also know that large datasets are sometimes misused. It is not to be panicky, but those are the risks that we need to mitigate. In my view, the single most appropriate way to mitigate them is simply to not disseminate that sensitive data into environments where you can no longer control what is happening to it. You need to make it available in TREs, where there are additional privacy mitigations, on top of the fact that people have to come to the data to work on it there. There also needs to be transparency about everything that happens to the data, first, because that reassures patients, the public and professionals that people are using the data only for the purpose for which they have been permissions, and secondly, because if you have a completely transparent research environment, you can quickly see if people are misusing the data.

There are various other techniques that people have tried to use on pseudonymised data to make it appropriate for dissemination and to mitigate the re-identification risk, but they all have big shortcomings. First, sometimes people try to minimise the data before they send it out. You might say, "The person who is requesting this data does not need to know the date, time and location of every diabetes event in that person's history. They just need to know whether that person has diabetes or not." You would then boil the data down and minimise it to just that before sending it out. That can be quite effective, but it also means that somewhere, someone behind the shroud is doing what the ABPI rightly says is 80% of the work, which is the curation. So that does not really fix the problem; it just pushes it further down the hill.

Another approach is people removing sensitive codes. By "sensitive codes", people mean that they will remove any traces of codes about mental health problems or reproductive health issues, for example. The problem with that is that my clinical work was in psychiatry, and I do not want to live in a world where we cannot do good research on mental health, and not just directly. We also want to know if mental health problems modify the relationship between other risks. Will they modify cardiovascular risks or your ability to engage with services, for example? Again, removing sensitive codes has serious shortcomings.

Then, there are other approaches, such as homomorphic encryption, but I can honestly say that I have not really seen a robust practical illustration of how they could be usefully implemented. Lastly, you can add random noise to the data. People hope that you can randomly perturb the data enough that you cannot identify an individual, but not so much that you destroy the true statistical relationships between the datapoints. The problem with that is that the more you perturb the data, the more you destroy the true relationships that you are looking for, and the less you perturb the data, the more the privacy risks persist.

Again, that comes down to how we have prioritised our spend on the use of data. We have thought of data as being like oil just sitting around waiting to be picked up and used. To my knowledge, UKRI and NIHR have never seriously invested in, for example, research that looks at mitigation and asks, "What's the right level of minimisation? If we minimise data in this or that way, how many unique matches persist? What are the challenges with some of the other mitigation approaches?" Thinking through the theoretical risks, applying them to repeatedly reused health datasets, and turning that into reusable code and methods that you could use in practice—that is the sort of thing that I think UKRI and NIHR should be resourcing for health data research in this country. But in the absence of that, it is just a list of things that other people have looked at a bit in other settings. It is not very clear that they are useful, and overall, they have a lot of shortcomings.

Q244 **Rebecca Long Bailey:** In much of the evidence that we have taken so far, people have raised issues about barriers across the UK data ecosystem. Your review made a number of helpful recommendations to overcome those barriers, and it comprises a package of reforms, but to what extent do you think they can be broken down into quick wins?

Professor Goldacre: I think that is really important. One of the biggest risks and shortcomings in the past has been a kind of destructive impatience. People have tried to do everything all in one go, and that often results in one single monolithic black box procurement, or multiple similar.

The recommendation in the review is that we build impatiently but incrementally, so that we can move forward with, for example, identifying three integrated care systems to adopt TREs and reproducible analytical pipelines; identifying three of the national cohort studies for research; identifying three of the most computationally mature national audit programmes; and identifying three national groups that are doing data analysis in the NHS for service improvement. You over-resource them—don't resource them to just get by, but recognise that you are building new working patterns and new methods that can be applied elsewhere, but also building capacity and capability as you go.

After two years of working with data pioneers in that way, using modern open methods alongside the current working practices, then I think you are in a position to have an explosion and transform the way that the whole country and the NHS service analyst community and the academic

community work with data. I think to try and do it all in one sudden massive burst would be problematic. In particular, that is because we need to build capability with people who understand both the software development and data science but also the specific challenges and opportunities in health data, in order for them to be a productive part of developing good code and good methods and good tools.

Q245 **Rebecca Long Bailey:** How long do you think it would take to embed data skills and the sharing of open code for data curation and analysis across the data ecosystem? Have we got a long way to go, or are we progressing quite nicely, in your view?

Professor Goldacre: There is better progress, encouragingly, in directly adjacent fields. As was covered in your previous work on open and reproducible science, the Office for National Statistics and Government Digital Service developed a set of best practices around open code for reproducibility and quality assurance, for other analytic functions in Government—the Government Economic Service, the Government Statistical Service and analysts in other Departments outside of Health and the NHS. They offered training. They have had very good progress in changing the way that people work. Similarly, adjacent parts of academia have embraced co-chairing much more.

There is a specific strategic shortcoming around health data research in the UK, which is that historically we haven't had a culture of that kind of data sharing—but, in particular among more junior people in that community, there is a lot of pent-up demand to work in computational ways. Again, that is why, if UKRI and NIHR offered high status, stand-alone funding for that kind of work, we would empower all the people who are already keen to adopt and work in those ways. It is a matter of a couple of years before you get a transformation across the whole system. That is because the old, closed ways of working are so duplicative, and they are so at risk of quality shortcomings, that they would be overtaken as soon as other competing methods were adequately resourced.

Q246 **Rebecca Long Bailey:** You have touched on this briefly already. Your review called for a frank conversation with the public regarding health data sharing, especially in relation to sharing data with commercial entities. What should that public engagement look like and what should the priorities for discussion be? Should it be an open consultation with the public or are there key priorities that you would like to see put in place?

Professor Goldacre: Obviously, I think patient and public involvement and engagement is incredibly important. I think it is not the only way that you earn trust. Historically, one of the problems is people have imagined, "If only we could tell patients the great stuff we do with data, then they would let us do what we want." I think we earn public trust by doing things that mitigate risks.

With patient and public involvement and engagement—PPIE, as it is known—my main concern on what I have seen historically is that you tend to get a very large number of quite superficial and duplicative projects all

looking at the same fundamental questions. A much better approach for the big monolithic recurring challenges that we have—commercial use, risk mitigation, how to make data accessible—is that it makes much more sense to do really big, systematic, robust pieces of work. I am a big fan of citizens' juries in particular.

There was a citizens' jury sponsored by NHSX and the National Data Guardian during the pandemic, which found that TREs were well understood and also strongly supported by the public. That involved days and days of hard labour, giving people information so that they could really dive deeply into the issues. I think that is the way we should do it. I am not so persuaded that the multiple duplicative and often quite brief PPIE projects that you see resourced at the moment are necessarily the way to address those big questions.

Q247 **Rebecca Long Bailey:** Thank you. I think it is understanding the different uses of data as well. One thing that strikes many of us is that people do not generally have a problem if they think that their data is being shared to develop new wonder drugs that will save people's lives, but what they do not think about is that that same data could be sent to a company that is developing fad diets—that data would still be used, theoretically, for research purposes, but for a very different purpose—and explaining that to the public is essential.

Given that, several witnesses suggested that there should be wider benefit to the public when data is shared with commercial entities. You suggested that the NHS should negotiate equity in innovations where the NHS is pivotal to development. What specifically did you have in mind? How would that work in practice?

Professor Goldacre: Look, I am not an economist; it is not my field. The Centre for Improving Data Collaboration in the NHS has been looking at this, and I am sure that it has good ideas on the tracks. Overall, if you want to drive innovation, you should as far as possible minimise the entry cost for people doing good new work with data, but also ensure that you get returns later where things prove to be effective. That is why I have suggested equity, but I am sure that better heads than mine thinking about intellectual property and how the state can foster a market and get returns from it would have better ideas than me.

Rebecca Long Bailey: Thank you, that is very helpful.

Q248 **Carol Monaghan:** Thank you, Professor Goldacre, for joining us this morning. We have been told about the ratchet effect with data, particularly the opting out of data, and how if a particular demographic opted out, it could bias the data, particularly health data. You recommend that a national opt-out should be reviewed when TREs are up and running. Might one solution be requiring periodic renewal of the person's opt-out?

Professor Goldacre: That is a reasonable suggestion. In my view, it would be unwise to revisit the shape or structure of opt-outs until we

prove to the public that we are working in a trustworthy way with data in trusted research environments. That is only time we can safely do it.

It is also correct to say that opt-outs can cause problems, and they are problems that are quite hard to foresee or predict. For example, it might not be a simple matter of demographics. In my group, we have done reports looking at quite fine-grained demographic and clinical details to see exactly which patients have and have not had the covid vaccine. It is possible that if patients opt out in large numbers, it might be that people who opt out of their data being used in that kind of analysis may also be the same kind of people who would opt out of having the covid vaccine, so certain types of analysis might be particularly vulnerable to opt-outs.

Because of that, in the future, if we are making data accessible through secure methods like TREs, a sensible thing to do might be, for example, to have an exceptions process, where you can say, "Okay, in general we run analyses on data where we respect patients' opt-outs," but if there is a very high-value or very important procedural or research question where you can make a good case for an exception on opt-outs, you could apply—through a standard set of rules formed through an open and accountable process—to override opt-outs in certain specific cases.

It is also perfectly possible, when working with data in TREs, to allow people to run an analysis first on a dataset that respects the opt-outs, and to publish those results, and separately, in principle, to run the analysis on the complete dataset that includes opted-out patients, and not to disclose the results, but to produce only a number, a coefficient, a statistic that tells you to what extent your results might have changed had you overridden opt-outs. That could be used to prioritise the situations in which you might make an exception on opt-outs.

I am therefore not sure about different ways of engaging with the public and patients in a labour-intensive way on opt-outs. It might be better to think strategically and fairly about the circumstances in which it could be reasonable to override opt-outs.

Q249 **Carol Monaghan:** We heard some interesting comments this morning from our witness from New Zealand, who talked about the possibility of an opt-in and that you could opt in for research only, for Government use or for use by commercial companies. Is that something we should be looking at?

Professor Goldacre: If I think through my own eagerness to engage with admin, I think people will not opt in. I think it will be a very unrepresentative sub-sample of the population. When I talk about overriding opt-outs, I do not say that lightly, but there are several practical things that the NHS needs to use to do with data where it is, I would say, unreasonable and impossible for the NHS to do its job if people are able to opt out of data being used. For example, you could not do QOF to look at payments to GPs that are based on whether or not patients with quite fine grained clinical and demographic characteristics have had quite a specific fine grain set of clinical activities. I do not think it is realistic to

ask the NHS to run that kind of service if it cannot access that kind of data. Similarly, it is very difficult to do service planning if 3 million out of 58 million people have opted out.

Q250 **Carol Monaghan:** Regardless of that, does the principle have some value? Rather than opting in or out, people would be able to decide, "I'm okay with my data being used for research but I don't want it being used by a commercial company."

Professor Goldacre: I have not strongly thought through different fine grained options around opt-outs. I suspect you would have fewer people opting out overall if you had more fine grained options around opt-outs, where you could say "I'm happy for it to be used for service improvement and for pure academic research, but not for commercial."

However, my worry is that the dividing line between commercial research and academic research is actually not as clear as one might imagine. For example, some of pharmacoepidemiology, looking at the risks of a given treatment for a given clinical outcome—an adverse event, for example—is done by academic researchers on public money, some is done by drug companies at the instruction of health regulators using their own resources, and some is done by instructing companies that they have to commission a study to do it and then they commission it from academia.

Often, the dividing lines between commercial and research are not quite as clear-cut. I think discriminating for data access on the basis of what type of person someone is does not quite do the trick, but I think you could possibly try to discriminate between different uses on purpose rather than person.

Q251 **Carol Monaghan:** Your main recommendations relate to a small network of trusted research environments. What barriers do you envisage in setting up such an environment and network, and what steps would you prioritise first?

Professor Goldacre: I think health suffers from being an early mover in working with data. If you look at the work of William Farr two centuries ago, his returns from individual hospitals are quite similar to the HES data that we still use today. Part of the problem with being an early adopter is that you have a lot of legacy projects and legacy infrastructure. That does not mean old copies of Windows 95 on computers; it means perhaps old teams and old working approaches that have a lot of status in the system and that are quite hard to segment out into the bits that are still useful and the bits that maybe are best left in the past.

One challenge is that it is a very crowded space with a lot of people who have been well resourced in the past to produce things a bit like TREs, but in many cases, as we were told by senior and junior stakeholders through the review, there is a sense that these projects have been resourced but not really delivered. There are things that we can do to improve the odds of delivery. One is to start small with pioneers, as set out in the review. The second, as set out in the review, is to have open and competitive funding for people who are producing the components of a good network

of efficient and secure TREs. I say that in particular because I think the previous approach to resourcing this has been to bypass the norms that you see in UKRI and NIHR, where you have open, competitive funding panels, where researchers present an idea and either get it funded or not. Those are capricious, arbitrary and infuriating, but they are probably the least bad solution for funding science. One of the reasons why British science punches so far above its weight is our funding structure, which is open and competitive. Those norms have been largely suspended when it comes to producing platforms for research, tools and systems and methods to join up the data and make it more accessible. For some reason, the choice has been made to resource those kinds of projects through special relationships or arrangements. There is a culture of projects being resourced, announced, talked up in blog posts or documents and then, as we were told by senior and junior stakeholders during the review, tending to disappear from view. Then another round begins.

I think all of that arises from the fact that we don't simply use the tried and tested routes of open and competitive resourcing. If UKRI, MRC and NIHR had an open and competitive funding panel, like they do for all other areas of science, to resource all of the questions we have been talking about here—data curation, secure analytics, risk mitigation in working with data—I think we would see an explosion of work. Of course, there is a role for the monolithic core funding for some projects and platforms, but what we have been missing is a rich, open, competitive and collaborative ecosystem in academia, producing good tools and methods, like we see in all other areas of science. That is why the UK has fallen behind. I think it is very reversible, but it requires UKRI and NIHR to take the simple step of having open, competitive funding panels for those questions.

Q252 **Carol Monaghan:** The Government have already set aside £200 million for establishing the TREs. Are you confident that that is going to go some way to setting up what you are talking about and that funds will be allocated in the manner in which you are talking? It would be easy for that £200 million to disappear on projects that perhaps are not going to do what you are recommending.

Professor Goldacre: At the moment, it is worth bearing in mind that that £200 million is not just for TREs. It also covers, for example, accelerating clinical trials—a separate but very important area—and other aspects of better use of NHS data. It is also the NHS resource on that, so there is separately UKRI and NIHR resource in this space, which I am concerned is not currently being made accessible through those open and competitive mechanisms. The NHS traditionally doesn't use those vehicles for resourcing activity. There is also space for core investment into the core computational architecture, for example, but what we are missing is the glue that goes between a raw computational environmental that contains the data and the data curation tools, the secure analytics and the risk mitigation when working with data. It is all of that work that I think UKRI and NIHR should rightly be resourcing and that I do not see them resourcing at the moment through open and competitive means.

Q253 **Carol Monaghan:** How much would that glue cost?

Professor Goldacre: I think all of the money that has previously been spent on that would have been enough if it had been given out through open and competitive mechanisms. Were the money that we know is earmarked for that kind of work—for example, through the UKRI DARE programme—to be available through open and competitive funding, that would probably be enough to make a very, very good start in this space.

Q254 **Carol Monaghan:** How is that money being allocated at the moment?

Professor Goldacre: It is currently under discussion within UKRI.

Q255 **Chair:** I have a couple of follow-up points. Is the £200 million that the Government set aside for establishing the trusted research environments not going through UKRI? Is it all separate?

Professor Goldacre: I don't know. I think nothing has been decided about that, and I don't know how it is being dispersed. In my mind, at any rate, it is somewhat separate from traditional academic funding. Bear in mind that this is not my field or what I do in my day job.

Q256 **Chair:** I understand that. It has very important implications, though. Have you had any discussions with Ottoline Leyser or Andrew Mackenzie at UKRI as to the need to follow the direction you have pointed out?

Professor Goldacre: Not since the review came out. You are right that we perhaps should reach out or write to them. I think that is a good idea. It is a peculiar position to be in—writing an independent review for Government. I wouldn't say I feel shy, but I will write a letter.

Q257 **Chair:** Good. Don't feel shy. We sometimes write letters to the Government, and I would encourage the practice.

Professor Goldacre: I think it would a tremendous help if you could write with a similar message.

Chair: We will be reporting and we are certain to pick up—I am sure my colleagues will agree—many of the recommendations that you have made here.

Q258 **Aaron Bell:** Thank you, Professor Goldacre. It has been a pleasure to listen to you over the past hour or so.

I will follow up on what you have said. Do you think they made a category error? When they thought of these TREs, did they think of them as some sort of national infrastructure that needed to be set up in a different way from how we usually approach innovation?

Professor Goldacre: The challenge is that they are both. In the review, we set out a way of thinking about TREs, which essentially have three components. The first is the governance and service wrapper, which is the sort of thing that is well within the scope of expertise. There is genuinely very strong expertise in Government around creating administrative systems that have policies and then implementing them in practice. Ideally, you want to have the same service wrapper for all trusted

research environments, so that people cannot go approval shopping. Also, if you have multiple different governance approaches in multiple different TREs or data analysis environments, as you have seen in recent history, that is a duplication of effort around governance. It is a lack of consistency, and it is very unhelpful.

The second thing that you need in a TRE is the basic raw underlying compute—that is an adequately performing database that just contains the raw data—and some kind of environment where essentially you can press a button and throw up a machine with adequate CPU, memory and disk storage to do some work.

The third set of work is the thing that I think is most challenging, at least historically: the specific methods, code and teams that can work in versatile and very raw computational environments in order to make sense of that data. In a field such as accountancy, because there are millions of customers for accountancy software around the world, there is no need to build any new accountancy software; you can just use what somebody else has made. When it comes to working with health data for research or service improvement, there may be some things that are useful from other countries and other settings, but everybody's data is wildly different. The data curation work that you do is so much more complex, but also so much more open to challenge, review, improvement, iteration and problems that you need that kind of work to be done in the open and in a strategic way. It is more on the model of the Government Digital Service, for example, where you have not so much a division between build and buy, but a division between whether you buy one enormous monolithic black box thing or have product managers and software developer bureaus coming into Government and working alongside people who have the main knowledge and expertise, whereby you are upskilling people in Government and NHS roles and have them as a collaborative part of the process of building this stuff.

We cannot absent ourselves from building that kind of data architecture, and we cannot hope that we can get it built by cutting a single cheque. We need to engage with it as a serious technical problem, like all the others that we see, both in medicine and in academia. We would not try to say, "All of renal medicine will be done with one cheque by one organisation in a black box." We say, "This is a serious technical speciality that needs engagement."

Q259 **Aaron Bell:** I want to go back to your earlier exchanges with Zarah Sultana about pseudonymisation and the basic difficulties of keeping the information while keeping the privacy. One thing that we have heard in this inquiry is the potential for synthetic data to solve that problem, given the advances in AI. It has to be done in a TRE, but you can create a new dataset of fake patients who essentially have the same characteristics. Rather than adding noise, is literally creating fake patient data a possible solution to the problem?

Professor Goldacre: Two thoughts on that. First of all, that contains two quite different concepts, with both flying under the title "synthetic data".

One of those is very similar to what was discussed earlier around adding noise to data. Essentially, if you try to produce synthetic data that has the same shape and structure as the real data, the more closely it replicates the real data, the more that the privacy risks in the underlying data persist. Moving between those two compromises is an irreducible challenge.

There is a second approach, which is what my own group's TRE, called OpenSAFELY, does, where you have randomly generated dummy data that has absolutely no relation to the real data except in its structure. In OpenSAFELY, you write your curation and analysis code using that dummy data, but you never try and use that dummy data to actually run your analysis. Instead, once your analysis and curation code is able to run to completion, you press a button and it gets pushed through into a secure environment and executes against real data and then only the answers fall out, but no researcher ever needs to enter an environment where they are interacting with real, granular underlying data.

Those are two quite different types of working that both fall under the rubric of synthetic data. They each have different strengths and weaknesses, and working on that is exactly the kind of thing where I think you need to see an explosion of open-source work, but also expertise. It is really about resourcing the intermediate work in working with data.

When we look at molecular biology, we don't say we are going to fund, prioritise and celebrate the work of the person who finally produces the molecule that treats a given type of cancer. We fund all of the intermediate work along the way. People will publish papers and get grants for work on reagents, on individual genes, and on individual aspects that make up the full journey to producing and evaluating a new molecule. We haven't seen that in health data and so we haven't had good productive work in health data.

Q260 **Aaron Bell:** To summarise, would you say that the statistical problem of accurate data versus privacy is unavoidable and insoluble to some extent, so the security is the only real way to address the issue?

Professor Goldacre: Fundamentally, the single most cost-effective and rapidly deliverable risk mitigation that allows us to access data securely and grant access to a huge number of people very quickly is to work in TREs. Some of those other mitigations have potential, but it has been very ad hoc how we have worked with them and thought about them, and if we made them the subject of serious strategic investment by funders, we might be able to bottom out where they could or could not be used.

Q261 **Aaron Bell:** On TREs, I agree with you—that is the obvious way forward. It is not a concept that is out there remotely with the public at all at the moment that these already exist in some settings, and you obviously want to see more of them rolled out. How do we get that awareness and build the trust?

Professor Goldacre: In my experience, the basic concept of TREs is very rapidly understood by patients, but we must also remember professionals

and people in the policy community are part of the general public. Also, with citizens' juries, it is very clear that the concept of trusted research environments is readily understood by patients and the public, and that people make the kinds of decisions that you would expect about their trustworthiness, so I feel confident that, were that to be explained to people, that would be straightforward.

Q262 **Aaron Bell:** There is a role for journalists here, presumably, being one yourself as well.

Professor Goldacre: Yes, but you have to remember that the public divides— I would say, from being a clinician more than a writer, that it is extremely rare to meet somebody who doesn't have the— I think everybody has the intellectual horsepower to understand these things, but people are very different in how motivated they are and how much they care. Actually, I think that most of the public don't care; they just want sensible decisions to be made on their behalf by people who are thoughtful, and they will use red flags like 3 million people opting out or a very successful campaign or credible stories about an absence of good risk mitigation as shortcuts. The concepts are readily understandable.

Again, first of all, it is not just privacy; it is also transparency. The thing about TREs is that where they are done well, they can prove to patients, the public and professionals that only permitted work is being done. When you download data on to your own laptop, nobody really knows what you do with it next. Secondly, they don't just bring privacy benefits. If we de-duplicate, reduce the unnecessary multiplicative spend from storing the data in multiple different places, store it in different ways so that the curation and analysis work is not portable between settings—the multiple governance structures that are set up in these multiple small projects—and harmonise around a small number of standard ways of storing and accessing the data so that everybody can reuse everybody else's curation, analysis and visualisation code, the productivity benefits will be enormous.

Lastly, we don't just want to produce faster or safer ways for current users of data to access data. Frankly, we also want to see an explosion of new, better ways of working with data. We want new entrants into the system. I have talked about the shortcomings of pseudonymisation. The biggest problem is not that there is a large number of frightening things happening out there; it is that because dissemination is so evidently unsafe, all the people working in information governance are rightly anxious, and therefore no matter how much you try to reform the information governance rules, the access will always be slow and it will always be a small number of people getting through the gate to get access to data. If we make it available through TREs, we will get more researchers, more service improvement and more life sciences innovation.

Q263 **Aaron Bell:** You have basically covered what I was going to ask. Your review also recommended streamlining governance arrangements. Briefly, what would you recommend, in terms of the governance of TREs, particularly on improving access?

Professor Goldacre: At the moment, as I said, governance is slow because people are rightly anxious about dissemination. If data is accessed only through TREs, and furthermore TREs with credible privacy preserving techniques in place and transparency, it is reasonable that it would very substantially reduce the information governance barriers to access because the privacy risks would be mitigated. However, there is a second category of approvals, which is about the purpose and ethics of the project. It would be possible to do a project that perfectly respected patients' privacy but was none the less racist or offensive in some way. You still need clear governance around the purpose, but you can substantially eradicate the unnecessary IG around privacy.

By doing that, you also crucially create a world in which you are drawing people towards more secure ways of working. That is an incentive to work in TREs, because if you want to download the data on to your own machine and work with it in a way that is inefficient and duplicative, tends towards closed ways of working and is less secure, you have to go through an enormous amount of IG work. Whereas if you want to work in a situation where you are sharing your code, accessing the data securely and efficiently reusing other people's code, you can get permission very quickly. That is one of many things that will rapidly drive a long-overdue shift towards TREs.

It is worth saying that for as long as I have been an adult and active in this space—I should also mention that I am the second generation of my family to work on epidemiology and data research infrastructure—I have seen Ministers and other key senior stakeholders saying, "The UK has amazingly powerful health data. It is the best in the world. We can capitalise on this. We should be driving research service improvement and life science innovation." We are being overtaken. It has happened, because other countries have become more competent in the way they manage access to their data. The only thing we have in our favour now is that we have a very large and ethnically diverse population, which means that the results from working with British data can be applicable around the world. This is probably our last chance. If we don't do it now, we should stop saying that British health data is tremendously powerful and just accept that we're not going to do it.

Q264 **Aaron Bell:** Finally from me, AI obviously has huge potential, but as we have already heard in this inquiry, it also has challenges because you need as big a dataset as possible, rather than trying to confine it to the specific variables you are looking at. What role could a TRE for AI play? In keeping with your other recommendations, how can what people are doing with the analytics there be transparent? How can you avoid biases, arbitrary decision making through the use of AI, and the potential ethical quandaries you have just raised about racism and so on?

Professor Goldacre: AI TREs are a completely different kettle of fish. In brief outline, the big challenge with, for example, random forest models is that you can't see how they are working. You can't see why the algorithm is adjudicating that a given person is at high risk or low risk of a given outcome. That's why they can be at risk of hidden bias. That can also be

why they present new and complex privacy risks. You can't see—a human can't review and understand the structure of a random forest model. Sometimes, you can find that a random forest model, in ways you hadn't anticipated, may have encoded the characteristics of individual patients. Therefore, that model can itself present a privacy risk.

In the review, we say that a lot of money has been spent on AI. AI TREs are a different kettle of fish and they should be separately resourced, and thought about separately on a separate budget line. I stand by that. The vast overwhelming majority of quick wins are in straight TREs for the 95% of use cases that we already have, which we are already overdue on delivering efficient infrastructure for.

I have ideas about how you could do AI TREs. You could have the random forest model developed in a secure environment, then execute on real patient data only within that secure environment, and then spit out the answer about one patient from that secure environment, so that the random forest model, which has a disclosivity risk, never needs to leave the trusted research environment. That is the stuff of five years of deep thought by 100 people.

Q265 **Aaron Bell:** Do you think the Government ought to commission a further piece of work on AI and AI TREs, maybe by yourself or somebody else?

Chair: Now that you have finished this, have you got time on your hands to think about AI?

Professor Goldacre: I think we should do the basics well. I think we should start with a practical set of three ICSs, three national teams, three birth cohorts, three national audit projects. We should start delivering TREs with reproducible analytic pipelines and teams that know about how to do reproducible open computational data science, taking a proper strategic approach to data curation in those environments.

We should do that alongside business as usual, so that there is no panic about doing it, and so that there is no sense of futility when it can't immediately be done perfectly in three months, or in a single cheque cut to a single monolithic contract. We should do that and, when we have cracked that, which is only two years' work, then let us look at AI TREs.

Chair: Thank you very much. Finally, Katherine Fletcher.

Q266 **Katherine Fletcher:** Thank you for your time; it is appreciated. I am going to ask a very future question and then a practical question. Earlier in your evidence, you used the classic Tony Blair example of how data records could have identified an individual. Obviously, the coming revolution of personalised DNA medicine, and having your DNA code on your health record, puts that on steroids, doesn't it? It is individual and unique. It is, "You went to the doctor's on Thursday morning," versus, "Your code at the end of this is GGT ACC." Those are different levels of specificity.

I can understand with existing computer models and the TRE models you

are proposing how we can reassure the public of security, but what happens if quantum computing gets going? When I give you my data record that says I was in hospital in 1978, that has got to survive innovations in technology as well, hasn't it? What thinking have you done there?

Professor Goldacre: I have not looked at quantum computing. We were not tasked to think particularly about genome data. There are TREs for genome data. In fact, interestingly, because structural genomics and molecular biology are fields adjacent to health data, where people have really prized and valued code sharing, there is much more advanced work in that space, around people having reusable code, reusable tools and platforms. There are TREs for genome data, but I'm not an expert.

When it comes to new techniques for being able to identify patients uniquely in data, again, if I understand the question correctly, a TRE would protect against those. First, you would be very cautious about how you gave access to somebody to execute those kinds of analyses on data. Secondly, you would get them to execute those analyses in a secure environment, where you could see what they were doing. Thirdly, if they did uniquely identify an individual by using those methods maliciously and outside the permissions they had, you would be able to see that they had done that, and you would be able to see the information that they had produced at the end of their work, in their output folder in a TRE, for example.

Q267 **Katherine Fletcher:** That is really helpful and brings me to my second question, which is on practicality. I know that this analogy does not bear close scrutiny, but it strikes me that you are effectively trying to create the Wikipedia of health data—a set of data published within a set of similar structures that requires a body of volunteers or qualified individuals to keep an eye on it. I think all colleagues here will know how painful it is to get the obviously "totally accurate" entries on our individual Wikipedia pages removed.

Who do you go to and say, "Hang on a minute, that's not true," or, "Hang on a minute, you've just identified me in that data."? Given that computing is going to accelerate, and health data is going to become DNA data, who is this cohort of brave and noble warriors who are going to make sure that TREs are not abused? If you were in our shoes, how would you explain that to the general public, who have demonstrated their concerns not just about the here and now, but about the future?

Professor Goldacre: It is the job of people who run TREs to provide an output checking service—for example, to manually check that the final output tables and graphs for the TRE—

Q268 **Katherine Fletcher:** Yes, but who is going to do it? Are we going to create a department that Zarah runs, with 200 people checking the data? Do you see what I mean?

Professor Goldacre: When you run a secure analytics platform, you have to have a service wrapper that has people who are employed to make

decisions about who gets in; you have to have people who are employed to check the outputs; and you have to have people who are employed to run them. None of this is free. It is basic data infrastructure for the nation, but I do not think it is any different in that regard from anything else that we pay for. We pay for hospital electronic health record systems to the tune of billions of pounds—certainly over time.

Katherine Fletcher: I am a biologist; I actually love what you are talking about, but I am just kicking the tyres a bit.

Professor Goldacre: On things such as data curation, for example, just because code is open, that does not mean it is produced for free. In fact, open code is fully compatible with a commercial model. I am proposing that people pay on a buy-out basis, and that commercial organisations or, more likely, the people in the NHS who are already doing that data curation work, do it in a slightly more structured way in a coherent library, where it is shared and available to everybody else.

In that regard, the parallel with Wikipedia that does not hold up is that Wikipedia uses volunteers. Of course, this stuff is done by people who are paid by the public purse, like we pay from the public purse at the moment for all the rather dispersed, less structured and very duplicative approaches for the same work done less well. However, the Wikipedia analogy is more sound on the structure. Wikipedia does not just say, "Here is an arbitrary Google document on the internet where anybody can write anything." There are structures within entries; there are standard stereotype structures for biographical things, for historical things; and there are also some invisible data structures such as pages, as a subset of the whole dataset, with structured data fields for age, nationality and so on. It is a matter of employing people for the operation—

Katherine Fletcher: Wikipedia has hundreds of thousands of volunteers globally. That is what worries me: the scale. Your answer is very comprehensive. I have one final—

Professor Goldacre: It is the work of dozens of people, not hundreds of thousands.

Q269 **Katherine Fletcher:** Oh, that is very helpful—I appreciate your expertise.

Finally, my colleague Rebecca Long Bailey said that people are willing to take an infinitesimally small risk with their data if they think it will help others, and I want to expand on that. The data has commercial value—commercial in the absolutely pecuniary sense—and would allow us to save people's lives by saying, for example, "You really need to give up smoking because your risk is so much higher." You described the data as the crown jewels of the NHS. Would you support a model in which companies pay for the NHS via paying for access to this stuff? Colleagues in this room could say to our constituents, "Please do not opt out. Here is why it is safe, and here is why it helps keep our NHS going." Is that a model that your work allows for?

Professor Goldacre: As I said, I'm not an economist, so I'm not an expert in thinking through the strengths and weaknesses of different ways of getting money back from the commercial sector. Also, as I understand it, in Government there is always some anxiety about having income for Government that is then strictly labelled for a particular class of expenditure, and that can have other second-round downsides.

In outline, however, I am personally very happy with the idea that if we can make data accessible in a way that is provably secure for patients' privacy, where we do that with commercial entities we should expect to get money back and we should expect to get money back for the state. We could label that, formally or informally, as helping to fund—

Katherine Fletcher: VAT was invented a century ago; we have probably got room for innovation.

Q270 **Chair:** A couple of final questions from me. Referring to your conversation with Aaron Bell about how the public could have confidence in trusted research environments, that was based on the experience of citizens' juries: when TREs were explained to them, people bought into them, as it were. However, the thing about citizens' juries is that, over a period of time, people have it explained to them what something is all about. That cannot be extended to the whole population; they will glean their information from news stories and snippets, which will often be about things that have gone wrong.

How confident are you that a way of doing things that can satisfy people's privacy concerns can be communicated in a way that makes it happen or makes it work?

Professor Goldacre: First, it is important to set this in the context of other pieces of civic infrastructure that manage risk. We don't expect that everybody in the whole population will understand how we manage fire safety in great detail. There is trust that it is done in a sensible and appropriate way; there is trust in institutions and in the methods. Also, in a transparent society, there is trust that, where something goes wrong, we will see it being reported in the media and we will see it being competently discussed in settings like Parliament.

On trusted research environments, first, there is communicating and explaining the fact that you are changing the way you work to a new, more efficient and safer way of working. Secondly, however, you are right to allude to the fact that people will judge these kinds of systems and platforms by the exceptional cases where they go wrong, and I think rightly so, because data, when it is leaked, cannot be unleaked; it is, in that regard, very similar to nuclear material.

That is why I think that each time we make a mis-step in this space—as we did with care.data and as we did with the attempt at the GP data extraction and dissemination in the middle of last year, each of which resulted in 1.5 million people opting out—we sacrifice public trust going forward. That is why it is so important that we stop making those mis-

steps. If we adopt TREs and if we stop pushing data out, I think we will have laid the firm foundations to not have disasters like that in the future.

The second recommendation—or perhaps the fifth out of 30 in the summary—is that we should map all current flows of GP data, because a lot of those flows are outside any central oversight, control or governance. In fact, there is no list of all of the different places where hundreds of practices have made a relationship with one entity that then aggregates the data. If we don't map those and understand them and think strategically about how to shut them down as soon as we can replace them with better, safer and more efficient ways of working with the data—if we are not careful, we may have a privacy problem with some of those in the future. That will undermine confidence as well.

The way to earn public trust is to do something concrete that earns public trust, and every time there is a problem, we sacrifice that public trust for a very long time; we lose it for many years.

Q271 **Chair:** This is my very last question. You may have heard in the evidence we took from New Zealand this morning that a new Bill before the Parliament there will make it a criminal offence to override the pseudonymisation—in other words, to identify an individual. Is that something that you think we should have here?

Professor Goldacre: Yes, very strongly. I think you should have barriers that make it hard for people to re-identify people—for example, people working in TREs—and you should have privacy preserving tools inside those TREs. I think you should maximise the chances of detection, and the best way of doing that is by having an audit of everything that happens in a TRE and, ideally, transparent logs of everything that happens in a TRE, so that if anybody did anything mischievous, it would be visible to everybody in the community.

Thirdly, you need huge penalties. People often point to GDPR; the problem with GDPR is that it is a very large financial penalty for an organisation. The penalties for individuals need to be absolutely enormous in order to communicate very clearly that this is not acceptable. I have certainly been surprised when I have looked at what has happened in cases where people have misused data. I looked at this in some detail, because you often find people from the epidemiology research community—quite relatably, I think—saying, "This all sounds like a bit of a bore. I quite like downloading data on to my research assistant's laptop and doing it there, and also no epidemiologist has ever misused data."

The problem with that is, first of all, there are very few audit trails of what happens in epidemiology. When we look at other datasets, you can see that it is actually fairly common for datasets to be misused, and the penalties are often quite slight. The two big examples one might look at are individual GP practices, where you will sometimes find somebody snooping on the medical records of a girl they used to go out with or somebody they were at school with, for example—the penalties for misusing that data are actually quite trivial—and, perhaps more

concerningly, a recent example from last year. There was a BBC news story where over 30 Metropolitan police staff—a mixture of officers and civilian Met staff—were allegedly caught and are under investigation for illegally accessing the detailed case notes of Sarah Everard, the woman who was kidnapped and murdered by a serving police officer. That is over 30 people working in very trusted roles illegally accessing data outside of the purposes of their work, even in an environment where most or all of them must have known that they were subject to audit.

So I absolutely agree: you need to block people misusing data, ensure that you detect it when they do, and make sure that the penalties are so high that they get talked about and people are really afraid.

Q272 **Chair:** You are talking about prison sentences, by implication.

Professor Goldacre: I do not think that would be unreasonable. When you talk to people who work in the MOJ or similar, or in the law, it is a bit like the school curriculum—everybody has a bee in their bonnet and everybody wants their thing to get the most attention—but yes, I think people who misuse data at scale or for malevolent purposes should expect very serious penalties. But we should also make it so hard to do it that the only people who do it are really trying hard to do it, so that you can say with absolute confidence, "This wasn't a slip. This wasn't somebody saying, 'I'll just have a quick look out of interest.' This was somebody who really had to go to enormous amounts of effort to maliciously misuse data." Then, it really is reasonable and proportionate to have very serious penalties.

Chair: Thank you very much. That goes to the purpose of this inquiry and, indeed, your work, which is to make sure that there is the trust there to allow us to gain the advantage of the enormous breakthroughs for millions of people that can come from research. That is jeopardised by lax practices, inadvertent or deliberate.

Thank you very much indeed for a very comprehensive and fascinating evidence session, Professor Goldacre. That concludes this meeting of the Committee.