HOUSE OF COMMONS

Science and Technology Committee

Oral evidence: Reproducibility and Research Integrity, HC 606

Wednesday 1 December 2021

Ordered by the House of Commons to be published on 1 December 2021.

Watch the meeting

Members present: Greg Clark (Chair); Aaron Bell; Dawn Butler; Katherine Fletcher; Mark Logan; Rebecca Long Bailey; Graham Stringer.

Questions 1 - 71

Witnesses

I: Professor Neil Ferguson OBE FMedSci, Professor of Mathematical Biology, Imperial College London.

II: Professor Marcus Munafò, Chair, UK Reproducibility Network Steering Group; and Professor Dorothy Bishop FBA FMedSci FRS, Professor of Developmental Neuropsychology, University of Oxford.

III: Dr Ivan Oranksy, Editor-in-Chief, Spectrum, and Co-founder, Retraction Watch; and Dr Janine Austin Clayton, Associate Director for Research on Women's Health and Director, Office of Research on Women's Health at the United States National Institute for Health.

Examination of witness

Witness: Professor Ferguson.

Chair: This is the first hearing of the Science and Technology Committee's new inquiry looking at the reproducibility of research and questions of research integrity. We will be taking further evidence during the weeks ahead.

I ask Members to declare any interests relevant to this session. I should declare that I am an honorary fellow of Nuffield College, Oxford.

Mark Logan: I am enrolled on a part-time PhD course in political science.

Q1 **Chair:** Thank you very much indeed. I have no further indications from other Members.

This session is to be an introduction to some of the questions that we will consider later in the inquiry. We are very fortunate to have Professor Neil Ferguson, who has helped the Committee hugely during the Covid pandemic. He was going to join one of the sessions later this morning, but, given that he has important meetings to attend on the Omicron variant, he has very kindly agreed to appear earlier. We are going to start with some specific questions before we then go back to the general.

Professor Ferguson, thank you very much indeed for appearing and for the work that you are doing.

Professor Ferguson: It is a pleasure.

Q2 **Chair:** I will start off with some initial questions and then hand over to my colleague Graham Stringer.

One of the reasons we are interested in your perspective is that one of the applications of reproducibility is in modelling. Obviously, modelling has been very prominent during the Covid pandemic, and the modelling team and work that you do at Imperial have been very prominent within that.

When you have a model that has been built up over many years, is it available for other researchers to replicate—to apply it to data independently to see whether they come up with the same or different results?

Professor Ferguson: I suppose I would step back and say there are two forms of reproducibility that are important, one of which is more important than the other.

The first is a narrow definition: you take a piece of computer code and you can run it yourself and get the same results. All the models that we have developed over the course of the pandemic fulfil that narrow criterion.

A broader aspect of reproducibility is whether two separate groups, taking the same assumptions—maybe input data, but with very different code—reproduce qualitatively the same results. It is the latter that is more important in science in general, and even experimental science, than the narrow one.

The UK Government have long experience of this. Particularly in a crisis, it demands the broader form of reproducibility, not relying on any single code base. All the groups inputting into SPI-M and SAGE make nearly all their code available in real time and have done throughout the epidemic. For instance, on critical decisions, there is input from the models that we have. We have multiple models of the Covid epidemic at Imperial College, the London School of Hygiene and Tropical Medicine—John Edmunds, whom you know—and multiple other groups across the UK. It is the congruence of results from those different, independently developed models that gives a degree of reassurance in the reproducibility of the science.

Q3 **Chair:** Thank you. On one aspect of that, which is to apply the same model to the same or similar datasets, one thing that one might want to do as a researcher is to vary some of the assumptions. Is it possible to do that with the code that you have published, or is it a package that is impervious to tweaking and has to be applied in toto?

Professor Ferguson: We have now four independent Covid models out there, from the most complex being the individual-based model we used early on to simpler models but with lots of parameters. In all those cases, if you download the source code, there are input files with parameters that are human-readable to some degree, and you can tweak them.

Mathematical models in any discipline codify a lot of scientific knowledge about systems. Not all changes of parameters make sense. You need the domain-specific knowledge to make intelligent choices about what things to look at rather than blindly tweak parameters.

On input data—you mentioned data—models are used in very different ways. During most of this pandemic, once we had data in the UK, we have been using models—simple statistical models to much more complex models—to understand the trends in the epidemic and to fit those models to data. That is an inferential process. There is a whole statistical engine behind that, and you get estimates of parameters out of the other end with uncertainties. One of the things about just downloading a model is that you largely miss that unless you go through that inferential step and you get a sense of where the uncertainty lies.

All our models throughout the pandemic—both ones whose final versions appear in scientific papers and ones that just go into SAGE briefing documents—are available on GitHub.

There is a slightly bigger issue with data availability. Generally, we have found UK HSA to be very open. There is a lot of pandemic data out there.

Generally, in our scientific publications, we have been able to make the data available as well. There are clearly some exceptions to that to which we can turn later.

Q4 **Chair:** Thank you. I am grateful for that. I have a final question before I turn to Graham Stringer.

I quite understand and appreciate your point about there being independently constructed models as a test alongside the existing models. In terms of the pure and narrow definition of reproducibility, you said you have published the code in journals and the datasets are available as a public resource. Are you aware of any substantial attempts literally to replicate the work that you have done at Imperial? One question that we are going to be pursuing is whether that is an interesting thing for researchers to do rather than discover new things. Has anyone felt motivated to do that?

Professor Ferguson: Multiple groups have looked at different models we have developed during this pandemic, some from a rather critical perspective and some from a more positive or neutral perspective. A group in Edinburgh, for instance, used a complex simulation code that we developed at the start of the pandemic and published a number of papers on that, making use of that code. There has been formal verification. A group in Cambridge took the code and reproduced the results exactly.

My colleague, Professor Samir Bhatt, developed something called "epidemia", which was used to estimate the effect of non-pharmaceutical interventions, but it can also be used to estimate that famous reproduction number over time in different geographies. That has been used across the world by groups. It is a relatively simple piece of code that is easy to reuse. Maybe that is something we can come on to.

The reusability of code is a key priority of the centre that I head. If code has broad applicability, like estimating R in an epidemic in different areas, we put greater investment into making it easy for other scientists to reuse. It is a priority for us. The whole discipline has moved in that direction in the last 10 years, particularly for simpler analyses, to packaging those sorts of analyses—there are modules in software like ARL or Python—that everybody can then reuse.

Chair: Thank you very much indeed. I am going to turn to Graham Stringer and then Aaron Bell.

Q5 **Graham Stringer:** Can I follow that point up briefly? When you publish your papers, is the code available at the time of publication?

Professor Ferguson: Generally, it is made available at the time of publication. It has evolved throughout the pandemic. Increasingly, all the code is available all the time for people to inspect. The challenge with that is that we cannot make the data available, and we have to make a snapshot of the data available with the publication. Generally, a good two thirds to three quarters of our code development that occurs using online

versioning systems such as GitHub can be viewed all the time. It is a continuous process.

Q6 **Graham Stringer:** At the time you publish both the code and the data, can other researchers replicate the work from the information that is available?

Professor Ferguson: If I think of all the peer-reviewed papers we have put out in the pandemic, I cannot be absolutely sure, but I would say yes in the vast majority of cases.

Q7 **Graham Stringer:** Thanks. Yesterday, in the debate on public health and the new restrictions that have been brought in, your work was criticised by one Member of Parliament, who basically said that, while there was independent verification of your work, predictions that had been made on foot and mouth, BSE, bird flu and swine flu were excessive partly because he claimed that workers at Lund University and John Ioannidis said that the assumptions you made were inflated. Do you have a response to the points Bob Seely made?

Professor Ferguson: That is something of a lockdown sceptic meme, and it has been fact-checked in a number of websites. It is mostly inaccurate. For instance, we did not really make excessive predictions for BSE. What we did was estimate the range of possible outcomes given the exposure of human population. Of course, given the uncertainties, you get a very wide range, but the central estimate was nowhere near as excessive as is often reported by some with a particular axe to grind.

Q8 **Graham Stringer:** When the predecessor Committee of this Committee interviewed Professor Jones of CRU at the University of East Anglia just over 10 years ago, I think he shocked the Committee by saying that when his work had been peer reviewed none of the peer reviewers had checked the basic input in terms of code and data; they had just accepted that and the methodology. Can you very briefly tell us, when your papers are peer reviewed, what work goes into it by the reviewer?

Professor Ferguson: It is a difficult question to answer for one's own papers because you do not see the work that goes in. In general, in the peer reviewing I do, it is quite rare in my experience to dive into code. Occasionally, one does it if you have a serious concern, but it is very time-consuming to go to that level of detail. That is true in reproducibility in science in general. There are some trade-offs between how much financial and personnel resource investment and time you want to put into that process versus getting more science out there, reviewing it effectively and critiquing it after it is published.

Q9 **Graham Stringer:** Can I move to a question of science philosophy, if that does not sound too pretentious? I am sure you are aware of Karl Popper's definition of science being falsifiability. How do you falsify something that comes out of a model?

Professor Ferguson: Models are used in very different ways. Most of the modelling we do is retrospective. It is analysing data and trying to interpret trends rather than making forward predictions. Modelling in that sense, just like statistical modelling analysis of any experimental data, is a tool for understanding processes and phenomena. Sometimes, you then come up with hypotheses about what is driving what, which can then, in some cases, be falsified. In other cases, they cannot. That is true of a lot of observational science.

The Popperian definition of valid science is quite a narrow one in some ways. There is a subset of modelling that is forward looking. That can be formal prediction or forecasting, and that is able to be validated, not in a yes/no sense, but in a quantification of the accuracy of that prediction.

Another form of modelling is used in economics and used sometimes in epidemiology, which is forward-looking counterfactual modelling. Here, we would be looking at what the possible trends are in, let us say, the epidemic and what the effect of different policy interventions might be. That is much harder intrinsically to falsify—it is also much easier to criticise, therefore—because reality is never going to match what is modelled. Policy is never exactly what is modelled. We cannot predict human behaviour and how a population responds that accurately. Things like variants will come up that we have not anticipated.

I would still argue that that modelling has value if it is based on the best scientific understanding of the phenomena at the time because, without it, policy makers are just relying on intuition to make policy judgments about what may happen in the future and what the effect of policies would be. I accept that, in a philosophical sense, it has challenges in terms of falsifiability.

Q10 **Graham Stringer:** You would agree with the famous quote from George Box that all models are wrong, but some are useful.

Professor Ferguson: Yes, indeed. I would distinguish. There are some statistical models that are just about quantifying the correlation between variables, and, in some sense, they are not wrong or right; they are just ways of summarising data. In terms of the transmission models, the mechanistic models that try to represent the epidemic process, they are all gross simplifications of human society and human interactions, by definition. We have learnt over many decades which of those simplifications seems to be a better description of the phenomena we see, but we never capture everything.

Q11 **Chair:** In one of your answers to Graham's question about peer review, you pointed out that peer review is done as a voluntary activity; it is not remunerated and it does not involve any publication, clearly, on the part of the reviewer. Your experience was that it is too consuming of resources to engage in very detailed work on checking assumptions.

Professor Ferguson: Sorry, can I cut in?

Chair: Yes.

Professor Ferguson: There is one exception to that, and I have acted in this role sometimes. Some, let us say, higher-profile journals, particularly the biomedical journals, will often have one specialist reviewer. Sometimes they pay for a statistical reviewer whose role is solely to evaluate the methodology used to analyse data. That does not typically mean that it will run code, but it means they put a lot more time into scrutinising whether the methods are the right ones to use and whether the results look plausible given the methods employed. I, and many of my colleagues, have acted in that role. Typically, it might take me three or four hours to do a peer review of a publication. If you are asked to do a detailed statistical review, it could take a whole day, for instance. Sometimes you will go back to the authors and demand more information and more analyses to check something.

Q12 **Chair:** Thank you. That is very helpful and very illuminating. If one thinks of other areas of life in which people rely on assumptions being sound and accurate but without the ability or the justification to check them—think of companies and audit statements, for example—there is a system that deals with that. You have auditors who are there precisely, often on a sample basis, to look at whether assumptions made are accurate. Do you have any reflections on whether the current system of peer review is good enough, or, perhaps in a subset of very consequential papers, should more resource be given and funded to be able to do that in a less voluntary way than happens at the moment?

Professor Ferguson: Stepping back—I can only talk about my discipline and related biomedical disciplines—I think there are reproducibility challenges, but I would not necessarily agree with the term "reproducibility crisis". In fact, in the last 10 years, we have been moving in the right direction, not the wrong direction. We are not in a perfect place in code availability, but we are in a much better place than we were 10 years ago.

As to your specific question, some of that happens already. The most important clinical trial papers get considerably more scrutiny than a much lower-impact paper going into a much lower-impact journal. That is why I talked about statistical reviewers. Those journals have those specialist individuals in place.

Could more be done? It could, but there will be trade-offs. For instance, let us talk about the AstraZeneca vaccine. There was enormous pressure to get out the results of that clinical trial. The trial was not a perfect trial. There were many challenges faced. Some of the interpretations and some of the data coming out of that trial were ambiguous. Should we have delayed it by a month or so for some independent group to process the same data to go back and forth, or is it better to get it out there with some limitations? In that case, it was better to get it out there.

I should say the Oxford group is also very open about sharing the data and open about the limitations. I am not picking on them. It is just a single example.

It depends on the urgency of the science. There would be a cost to moving to a more audit-type model. There is a cost to all efforts to improve reproducibility. Scientists, particularly in academic environments, but even in commercial environments, are under a lot of pressure to deliver the maximum science per pound they can possibly deliver. Governments also tend to rate scientific productivity on those same metrics. There is a balancing act. I do not have a particular view on where to draw that line. I am optimistic. I think we are going in the right direction in code and data being available.

Chair: Thank you. Cost and speed are the two consequences of taking more time and having more eyes looking at these things.

Q13 **Aaron Bell:** Thank you, Chair, Professor Ferguson, for coming once again to our Committee.

Widening it out, as you just did, do you think that the pressures that end up being placed on researchers by the academic system and the publishing system compromise reproducibility in a general sense because of the pressures that people have to get published?

Professor Ferguson: To a degree. The incentives, certainly until recently, have not been there to allow easy reuse, for instance, of code and data. Most scientists collect data—we are unusual in not doing very much of that—whether in experiments, observational studies or clinical studies, and then analyse that data. We are getting much better—and the pandemic has accelerated that—at then putting out the data and putting out the code. The data has been curated quite quickly and the code has been written typically for a single set of analyses without much thought about whether we could generalise this and make it easy for people to reuse. We would need both to invest more money and change the incentive system to move out of that domain where scientists are rewarded for making it easy for people to reuse their code and data and extend the work.

There are plenty of scientists who are motivated in that direction, including in my discipline. There are lots of young scientists coming through whose first thought is about how they can package the analysis they have come up with in a way that enables easy reproducibility and reuse, more importantly, but it is not necessarily the norm yet.

Q14 **Aaron Bell:** In written evidence that we received, there is a number of questionable research practices or elements of poor study design. To briefly list them, there is what is called HARKing, which is generating hypotheses fitting the results; p-hacking, where you manipulate results until you get something that is statistically significant; outcome switching, where you report only favoured results and not other ones;

and various other processes like trying to collect more data after you have some results to get yourself to that p-hacking point or whatever. Have you seen any of that in your working life? I do not mean in your own work; I am not suggesting that. Have you seen that in work that you have peer reviewed or work of colleagues?

Professor Ferguson: Yes, undoubtedly. There are statistical methods for mitigating that. What you are talking about is observational studies or basically post hoc hypothesising about data and generating lots of hypotheses and seeing which one is statistically significant. It is questionable behaviour, but we know how to stop it. The American Statistical Association has said this already; moving away from the obsession with P values and statistical significance would be a good start.

There are also rigorous statistical approaches—Bonferroni correction, for instance—for allowing for multiple testing. There is a whole statistical literature on exactly that, of data mining techniques, and trying to judge, if you do 100 tests on a complicated dataset, how you work out that you have found a statistically significant result just by chance or whether it is more plausible. Those methods are often used, for instance, in bioinformatics disciplines in genetics, but they are taking a while to move across the whole of biomedical science.

Q15 **Aaron Bell:** How rare is outright fraud? Some Covid papers have had data that has been made up when people have looked at it. We have gone through a period where we have been seeing a lot of preprints rather than peer-reviewed papers. How rare is outright fraud in the data that people are using for their research projects and their papers?

Professor Ferguson: I do not think I am the best person to answer that question. I do not think I have encountered this in my career. I have certainly encountered plenty of papers that are fundamentally flawed and have made a fundamental misjudgment about how to analyse data or interpret it or construct a model or whatever. I cannot think of an instance. I think it is relatively rare, therefore. I have not encountered it or, let us say, I have not picked up on it even if I have encountered it. There are clearly plenty of high-profile instances where it occurs and gets reported. I have sat on a disciplinary panel at Imperial, which I will not go into detail about, where questions of data integrity came up. It happens, but it is a rarity.

Q16 **Aaron Bell:** The bigger issue is the questionable research practices and the incentives within the system.

Professor Ferguson: Yes. Regarding questionable research practices, people like to think they are good. Very few people have a malign intent in doing this. They are excited by their work. They are excited by the data they have collected, and they clearly want to find something important in it. The role of peer review and guidelines and scientific integrity rules is to try to mitigate that enthusiasm.

Q17 **Aaron Bell:** Things like registering hypotheses in advance would be an

issue.

Professor Ferguson: Yes, that is one thing. Randomised clinical trials have to be registered in advance. A lot of the data we collected during the Covid pandemic is not from clinical trials; it is from observational studies. Those are not ever reproducible. You are observing one population at one point in time. There, the transparency and what data was collected but also the review and the statistical rigour of the methods used to analyse that data is key. You are always going to end up with ambiguity there. That is where there is a trade-off. Often, it is better for one group to publish something with a hypothesis and then somebody like me and other colleagues might say, "Actually, no, I disagree with that." That is part of the scientific debate. It is not a lack of integrity.

Q18 **Chair:** Thank you. Before I go to Dawn Butler, I want to probe a bit on Aaron's line of questioning on research rules and data mining. Are these rules not for a different age? In the world in which we have artificial intelligence and vast analytical resources, does AI not look into data and find connections that might have evaded public consciousness or recognition for which people do not have a theory in advance, but in throwing them up can make some very important connections for understanding and, in clinical matters, for improving people's lives? Are we in danger of constraining what might be a rich seam of discovery?

Professor Ferguson: It is a difficult one with registration. For instance, when you put in an ethics application—and that probably is the best analogy—you will say, "This is what we are going to do. We are going to collect this data and we are going to look for these things." Typically, it is not like a clinical trial registry. You give a general scope of work of the sort of things you are going to examine and the amount of data you are going to collect. I do not think that would constrain that sort of exploratory analysis. There is nothing wrong with writing when you register your study that we are going to perform exploratory analyses using AI methods to explore connections and, probably in that case, out of sample cross-validation to test whether those correlations you find are predictive. I do not think it necessarily needs to be constraining. You have a different goal from clinical trials, where you have a very specific endpoint that you randomise the study population for and very specific consequences for a lot of clinical trials in the registration of, for instance, new drugs and new vaccines. That is much more tightly constrained.

For general science, transparency about the types of studies that you are planning would not be a bad thing. In some sense, it happens already with grant applications. When scientists apply for research funding, for instance, in my discipline, they will say, "We intend to run a population cohort study that collects these variables. We will test these hypotheses. We will analyse this data." It is not giving code in advance, but it is saying what the aim of the research is. One could have greater transparency up front that can then be checked back when the peer-reviewed papers come out of that research.

Chair: I see. Thank you very much.

Q19 **Dawn Butler:** Thank you for your time today. What are the top three things that need to be implemented to build confidence around the work that you are doing?

Professor Ferguson: I hate top three questions. Is that work that my group is doing specifically, or the broad area?

Q20 **Dawn Butler:** The broad area of reproducibility.

Professor Ferguson: As I said, I am an optimist in this area because I think the trends are in the right direction, but more can be done, except where individual confidentiality is at stake, in demanding that data is released for papers. Some journals do that now—many of the higher-profile journals—but it is not standard practice across the whole of science. It would aid reproducibility, but, more than that, as everybody testifying to your Committee will say, it will aid the value we get out of the public money invested into research.

That would be No. 1. We employ—and this has been built up over the last 10 years—eight professional research software engineers in the centre I head. We are a large centre, but that is really quite unusual. Particularly for code that is going to be often reused, that is to improve software development practices. That costs money. There is a plea for research funders to prioritise that professionalism around software development just as they prioritised professionalism around statistics, for instance.

Thirdly, there should be a broad look at the incentive structures that we have already talked about. You should be able to be a successful scientist if most of what you do is deliver, for instance, analytical tools that help the general scientific community and move the discipline forward in that way, rather than getting high-profile, individual scientific insights that get into the top journals. I am sure you will hear much more about that later as well.

Dawn Butler: Great. Thank you. Happy birthday.

Professor Ferguson: Thank you.

Q21 **Chair:** In terms of reproducibility, you said that there is the code and the data. One can expect data that is gathered from publicly funded sources to be available, but in medical science that data has a very high proprietary value around it. To what extent is the current practice good enough in making that available to researchers beyond the ones that have been funded perhaps by a medical research company or pharmaceutical company to investigate the efficacy of a particular drug?

Professor Ferguson: Being slightly diplomatic here, considerably more could be done to release clinical trial data in general, whether publicly funded or commercially funded. It is not universal, but there are plenty of instances where pharma companies do not publicly release detailed

clinical trial data and individual level data because there are concerns that somebody might find something that would have a commercial impact on their product. There have been some quite high-profile instances, for instance, around Oseltamivir, the Tamiflu drug, where there has been toing and froing and a fight, and only the trial data has been released. It has harmed the company in that instance. Transparency on the commercial side is to the benefit of those companies and the development of next-generation compounds and vaccines. There is intrinsic conservativism within the sector that opening up data like that opens them up to additional potential risk.

Q22 **Chair:** If that is a concern and a problem, what is the solution? Is it moral suasion, is it research ethics, or is it compulsion?

Professor Ferguson: For the pharma sector, a lot of data gets released automatically through the regulatory process. Journals could do more in, basically, demanding, even if it clearly would be anonymised, that at least a minimal dataset required to reproduce the results in an important clinical trial paper of a new product is released with the paper. Journals like the *Lancet* do not demand that at the current time.

Q23 **Chair:** Thank you very much indeed. You have been very kind in talking to us about research reproducibility, but I want to be opportunistic and ask you a question about Covid, which, as you know, has been a great source of interest to this Committee. This Committee has heard from Sir Andrew Pollard, who was one of the discoverers of the Oxford vaccine, that the likely course of Covid, as with other coronaviruses, is that they will become progressively more infectious and they will evade vaccines in transmission, but less prone to severe illness. Does your experience as an epidemiologist conform to Sir Andrew's view of the world?

Professor Ferguson: Certainly for the former. Viruses will evolve to become more transmissible. That is their measure of fitness—how successfully they pass from person to person. They can do that by becoming intrinsically more transmissible—more infectious—like the Alpha variant was compared with the original strains and like the Delta variant is more infectious, or in a highly immune population they can gain transmissibility by evading immunity. We have not had it proven yet, but it looks like the Omicron variant may conform to that latter type of evolution. Influenza viruses do that all the time. The major way they evolve is by evading host immunity.

The latter hypothesis—that viruses evolve to become less lethal—is less supported by data. It happens in some cases. We lived with the smallpox virus for many centuries until it was eliminated, and it still killed a third of the people it infected when we eliminated it. In some sense, viruses do not care about whether they are lethal or not. If being lethal helps their transmission, they will be lethal. If it does not, their severity may wane over time.

Transmission is the selection pressure for viruses. It is too early to say whether Omicron will be more or less severe than previous variants, but what we have seen so far is that Alpha has been a little more severe than the previous strain in severe outcomes, counterbalanced by the fact that we have treatments, and Delta was more severe again. The trend we have seen so far is towards greater severity, not lesser severity, thankfully countered by better treatment, knowing how to treat people, monoclonal antibodies, antivirals and all the other drugs, which mean that people have a much better chance of surviving severe Covid today than they did at the outset of the pandemic.

Q24 **Chair:** I see. Does lethality not impede transmission?

Professor Ferguson: Only if individuals die before they transmit. If we look at what causes people to die from Covid, it is principally the hyperactive immune response. Most of the transmission has already happened by the time people get hospitalised. The virus cares about replicating very fast within the respiratory tract and getting out into the environment. If that happens to kill somebody 10 days later, the virus really does not care.

Q25 **Chair:** From your experience, how long do you think it will take to know enough about the transmissibility and lethality of this new variant to be able to get to the next stage of decisions?

Professor Ferguson: The first data coming out is experimental neutralising antibody titre data that says in a lab system—an in vitro system—how much more difficult it is for antibodies to neutralise this virus. We have learnt a lot over the pandemic, and that will give us an early indication of how much of a threat it is. That data, hopefully, will become available in the next week or two.

The second will be epidemiological data and looking at how this virus is transmitting in South Africa, the UK and other European countries where surveillance is more intense. That probably will take three to four weeks before we get that early indication like we did with Delta and Alpha of how much faster this virus is transmitting and how much it is compromising real-world vaccine effectiveness. We have to be patient. It is likely to be towards the end of this month where we have a clearer picture rather than in the next week or so.

Q26 **Chair:** Professor Ferguson, thank you very much indeed for helping the Committee with its new inquiry into research reproducibility and indulging us with some questions on Covid. You have been very generous with your time over the course of the pandemic with the Committee. On your birthday, I think we should release you, probably not for a day of festivities but for meetings on Covid. Thank you very much indeed for appearing.

Examination of witnesses

Witnesses: Professor Munafò and Professor Bishop.

Q27 **Chair:** I invite our next panel of witnesses to join us at the table. As they do so, let me introduce Professor Marcus Munafò, who is the professor of biological psychology and the Medical Research Council investigator at the University of Bristol, and Professor Dorothy Bishop, who is professor of developmental neuropsychology at the University of Oxford. Thank you very much indeed for joining us.

We are going back to the beginning, as it were. We went into some specifics with Professor Ferguson. This is the first session of our inquiry. Professor Bishop, how would you define "reproducibility" in this research context?

Professor Bishop: It is a word, as Professor Ferguson said, that is used rather differently by different people. The standard definition that is emerging that most people would accept is that it is the ability, if you have a set of data, to reproduce the results from that dataset that somebody else had reported.

Q28 **Chair:** How does that differ from replicability?

Professor Bishop: Replicability might be that you would have another study on the same question but with a different dataset but where you should be able to essentially get a comparable result.

Q29 **Chair:** Thank you. Before I turn to Rebecca Long Bailey, I will ask you the same question, Professor Munafò. Is that you how think of research reproducibility?

Professor Munafò: Yes, I would agree. All I would say is that to some extent we need to step back from that and think about why those things matter. They matter because we are fundamentally interested in whether the knowledge that we generate is robust and will be useful, and whether we are getting the right answer, which is not a trivial thing in the context of science. Those are ways of evaluating the extent to which we are getting towards the truth, the right answer and able to make progress.

Chair: Thank you very much indeed.

Q30 **Rebecca Long Bailey:** In 2016, a survey conducted by *Nature* found that 52% of researchers believed that there was a significant crisis in reproducibility. If we ran that survey again today, would we find the same results? Is there a crisis in reproducibility?

Professor Munafò: I do not like the crisis narrative, because it implies an acute problem that we can fix and move on from. I think there are problems, and I do not think the results of the survey would be radically different today, although a lot of progress has been made, but progress takes time. We need to reflect on how we can improve the way in which science works and take into account the fact that science is a human process and is going to be subject to error and various biases that are just the result of the fact that the work is done by humans. We need to

think about the extent to which we can focus more on the process than on the outputs and embed quality control into that, ensuring that the incentives that shape the behaviour of scientists are well aligned so that what is good for scientists' careers is good for the advancement of science as well.

Professor Bishop: I broadly agree. We might find differences between different disciplines. There has been a lot of movement in my discipline of psychology. There would be a view that things have improved, albeit from a fairly parlous state of affairs. There are other disciplines that have been rather slow to pick up and sometimes regarded themselves as rather superior to things like psychology and just assumed that the problems are restricted to psychology, where instead people are uncovering quite severe problems.

I take as an example cancer biology, where people try to do studies reproducing some fundamental studies in cancer biology. Whereas in psychology the difficulty was that when they tried to replicate them they found that a proportion of them did not get the same result and did not replicate.

In cancer biology, when they tried to do these studies, they could not work out how to try to do the same study because the descriptions of what had been done was not sufficient, so the protocols were not adequately clarified or the whole study depended on a set of reagents that only one lab had. It was really hard to get past first base. This study had aimed to try to replicate a whole set of cancer biology studies and ended up only being able to do that for a smaller number because for such a high proportion they could not work out what had been done. That is pretty serious.

It varies from discipline to discipline, but most disciplines are affected one way or another. As Professor Ferguson said, in some regards, things are moving quite fast. Over the past 10 years, a lot of positive change has occurred.

Q31 **Rebecca Long Bailey:** Thank you, Professor Bishop. What positive changes would you refer to when you make that statement?

Professor Bishop: In part, the sort of thing that Professor Ferguson mentioned about things being more open and transparent. When I am asked to peer review a paper these days, I usually refuse to do it unless the code and data are available. That is not standard. Editors often say, "Who do you think you are wanting code and data?" As far as I am concerned, if they want a proper job done, I need to have access to code and data. That is still not the default, but it is becoming more common, particularly in my area of psychology and in some areas of computational biology. That is one change.

Another change is that funders are realising that unless this problem, in so far as there is a problem, is fixed they are not getting bang for their

buck. You can have a lot of studies where it is just not very interesting, or where somebody has a hypothesis that is not confirmed. That goes with the territory. We call it research because we do not know what the answer will quite be, and quite often the answer is boring or not as nice as we thought it would be. The real problem is when research is either inconclusive or even misleading—where people are putting out these things that have come out of data dredging that are really just false positives, and other people then try to build on that. We have more awareness of that.

One of the things that really encourages me is the extent to which this has a grassroots element to it. I did not read all the responses to your request for evidence, but there were a lot, I noticed, by quite young people. Early-career researchers are very concerned about this issue because it affects them directly. When I talk about it, what is still disturbing is that, if I talk about how to do things properly, a very common response is, "Well, I would like to do that, but my boss will not let me because we have to get this paper out."

Another response is, "What do I do with that situation?" Sometimes, you end up with the best people deciding that they are going to leave rather than stay in the field because they do not really like to work in a way that is not really as kosher as it should be.

This is both bad and good. The good thing is that there has been a lot of grassroots input into bringing about change, not just in this country—although very notably in this country—but also across the world. Therefore, this is why we might start to see change. Unless we can really foster that, we risk losing the best of our young scientists, who are very impatient with the current system.

Q32 **Rebecca Long Bailey:** Thank you. Professor Munafò, what developments have you seen in recent years that improve reproducibility?

Professor Munafò: There have been a number. The starting point is that there is grassroots interest in doing things better. You are seeing a lot of energy coming from early-career researchers but also established senior researchers wanting to reflect on how we can improve the way in which we do science and put in place practices and processes that can support that.

Transparency is a big part of that. You are seeing much more emphasis on open research. That is something that researchers themselves have a great deal of agency around. They can adopt those practices, notwithstanding the fact that they may experience resistance from some quarters if, for example, their PI—their boss—is resistant to that.

You are also seeing that supported by funders and by publishers. We need to focus on how we can support that through co-ordination and ensuring that the way in which people work is interoperable—what the

funder requires is supported by what the journal requires and is supported by the training that is given to the researchers.

We can still work on that kind of co-ordination. The UK particularly has some natural advantages there because we have quite a well-connected and compact research ecosystem that allows us to do that better than many other countries.

We have set up the UK Reproducibility Network, for example, which is intended to bring together those different stakeholders and actors in the research ecosystem to foster that sense of collaboration and co-ordination. There is a very strong sense of competitiveness in science at an individual level and at an institutional level. That is not inherently a bad thing, but there are times when collaboration can move us further forward. There are plenty of elements of how we do research that are effectively pre-competitive where that kind of collaboration and co-ordination and that kind of sharing of effective practice would be simply more efficient. We are starting to see lots of that.

We are also starting to see lots of innovation in how we do science. We have technology available to us now that was not available, or at least available at scale, even 10 years ago that allows us to share data, code, materials and many other aspects of the research process so that we can recognise the intermediate research artefacts that contribute to the end product, allowing us to focus more on the process and less on the end product. By making that process transparent, we can better recognise individual granular contributions of individual scientists—this person developed the code, this person collected the data and so on.

That transparency can also serve as a quality control process so that there is more scope for scrutiny of those intermediate artefacts of the process itself rather than just peer review of the final paper. The fact that one might have one's data, code, or whatever it might be, checked leads to an incentive for the researchers themselves to make sure that their processes are robust and that the quality of what they finally put out there is sufficient.

The analogy that I sometimes use—and it was one that was made by Robi Blumenstein, who is the CEO of a charitable funder in the US—is that research is a bit like the US automobile industry in the 1970s: it is very much focused on productivity with lots of stuff rolling off the production line. If you are of a certain vintage, you will know the phrase a "lemon"—an irredeemably badly built car that however many times you fixed it would keep breaking down. The problem was that there was no quality control in the process that led to those outputs.

William Edwards Deming, the statistician, took the concept of quality control in manufacturing to the Japanese automobile industry. That is an industry that still has a reputation for reliability today. Manufacturing has been transformed. Cars are orders of magnitude more reliable now. The less intuitive insight that Deming had was that if you focus on quality

control during the process you, of course, improve quality of the output, but you also improve productivity because you are not having to expend resources on fixing cars that have broken down later.

If we can focus on quality of the scientific process, we should improve the quality of our outputs. If that analogy holds—and we would have to test this, but there are good reasons to believe it—we can improve productivity in the sense of generating knowledge that is robust, that advances our understanding of nature, and that leads to societal benefit to health impact and so on. A fundamental thing that is changing is that we are focusing much more on the process, using transparency as a means to do that, in part at least, and increasingly less on just the outputs.

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

Q33 **Graham Stringer:** Can I follow up directly on Professor Bishop's answer? We have heard in some of the evidence this morning that the reproducibility problem seems to me to be increasingly a problem of peer review. Do you think that it is a failure of peer review that we are getting some of the problems in reproducibility?

Professor Bishop: No, not really. I do not think that is a major factor. There are problems with peer review, but some of the bigger issues with reproducibility are in the researchers themselves. In my field, data dredging, or p-hacking as it is called, has been endemic. It leads to false positive findings getting into the literature, and the problem has been that researchers have genuinely not understood how big a difference it can make and how detrimental it is to science.

One of the people writing about this said that people tend to think it is like jaywalking when, in fact, it is more like robbing a bank. It is a serious problem for science. The statistical methods that people are using to evaluate their results lead them to conclude that they have something really meaningful when they have looked at a whole host of potential results from a paper. Professor Ferguson mentioned there's corrective methods you can apply to deal with this, but people do not do that.

I see that as the principal problem. Peer review is flawed. Some people say it is the best thing that we have. The real problem with peer review, to my mind, is that it comes at the wrong time. You might have noticed that quite a lot of people, in responding to your call for evidence, have mentioned pre-registration, or this thing of registered reports, which I am a very big fan of. The idea is that you pre-register what you plan to do. You write a very detailed protocol of what you plan to do, including an analysis plan, and that is what is peer reviewed. Then, if you do what you said you were going to do, it gets published regardless of whether you found something significant or non-significant. That gives you much better peer review because the peer reviewers can be constructive, and they can improve your research at that point.

At the moment, you send in a paper—all scientists have had this—and a peer reviewer comes back and says, "You should have done this." You think, "It's a bit late now. I've spent three years gathering my dataset." You can have a very hard time getting registered reports accepted because they are very rigorous in how they want you to do things, but it means that you then get the peer review at a sensible point in time.

I am very much in favour of that approach rather than the current approach, which is rather like locking the stable door after the horse has bolted. Peer reviewers are much happier to put time in if they feel that they can influence what you might do rather than just having to tell you what is wrong about what you have already done.

Q34 **Graham Stringer:** That is really interesting. To take the example you gave in your answer to Becky about cancer biology, should the fact that those experiments could not be carried out because the information was not there in the papers not have been picked up by the peer reviewer?

Professor Bishop: Probably, but there is often limited space for methods. You often do not realise until you try yourself to do something again what is missing. There is often a lot of technical detail. I am not in the field of cancer biology, but I know from my own field that when we have tried to do this with student projects, where we think, "Just reproduce somebody else's study," you think all the information is there, and only when you try to go through and see exactly what you would need to do you realise what is missing. Certainly, you could say that the peer reviewers should have picked up on something. The beauty of that is, if those studies had been pre-registered, you would have found much more detail being required in methods and exactly what they planned to do and how they planned to do it.

Q35 **Aaron Bell:** Going back to your original distinction, Professor Bishop, between reproducibility and replicability, it seems from what you have just said to Graham that the reproducibility stuff is about the methods and making clear what you are doing. If things cannot be replicated, that is saying that, essentially, there was something wrong with what you were doing.

Professor Bishop: No, not necessarily. What, in effect, you are trying to do in science is pick out a signal from noise. You have all this variation that is for causes that you are not interested in at all, and then on top of that you want to see if there is a real effect. There will be instances where somebody picks up on something and has done a perfectly valid experiment but somebody else is not going to get the same result. It is wrong to equate lack of replicability with there being something wrong.

What you would want to do before taking that work forward and building on it is to replicate it. You do not really want to go ahead—and the drug companies have been particularly concerned about this—and build on something and invest a lot of money developing a drug on the back of

some biomedical study if it turns out that that initial finding was indeed a false positive.

Q36 **Aaron Bell:** We have just heard from Neil Ferguson that we probably want to move away from p values, but, yes, of course, 5% of studies will be noise that is incorrectly interpreted as signal. If we are getting a lot more than that that cannot be replicated, that surely suggests there is something wrong.

Professor Bishop: Yes, but you cannot identify which of them is which. You cannot say of an individual study that there was something wrong. I certainly think there is something wrong with the literature. One of the things that is wrong is down to publication bias, in that people will often not publish their null results. That is another thing that would be solved by registered reports, because then you have pre-registered what you are going to do, and it will be published even if it is null.

There is a belief that finding out what does not work is not interesting, whereas, of course, it is extremely interesting. In a pandemic, you want to know which treatments do not work so that we do not start putting people on ineffective regimes. Somehow, that has always been regarded as less exciting than finding things that work. It has been well documented, if you follow through which registered clinical trials get published, about half of the ones with null results never see the light of day, whereas virtually all the ones that have positive results do.

Q37 **Aaron Bell:** This is not a new problem. Particularly in the field of psychology, we have not been able to replicate some of the foundational studies from the ’60s and ’70s.

Professor Bishop: Yes.

Q38 **Aaron Bell:** We have built an amount of literature off the back of that. That is something that, clearly, the field needs to address.

Professor Bishop: Yes. What has been somehow lost in all this is that science should be cumulative. If you want it to be cumulative, it is very dangerous just to take a single study and then develop more and more on that without first being absolutely sure that that effect is solid. Unfortunately, that has often happened, and that is because people typically have not been funded to do replications or have not been motivated to do replications.

There has been a suggestion that it would be a lot more sensible if the first project of people doing even undergraduate degrees, and certainly master’s degrees, was to replicate something else in the literature, whereas at the moment the general idea is that everybody has to do something novel and exciting. It is a bit ridiculous because it has been argued that this teaches them how to be bad researchers. They do not know anything, and they are supposed to go out there and discover something amazing and get better marks on their dissertation if they find something original. It would be much better training for them to try to

replicate an established finding on which other people wanted to build. That would be helpful to all of us.

Aaron Bell: Thank you, professor. Thank you, Chair.

Q39 **Chair:** Thank you very much. Professor Munafò, do you have any views on what we have heard from Professor Bishop?

Professor Munafò: A few things. I really agree with that. We need to recognise that replicability and reproducibility are not ends in themselves. There are broadly two reasons why one might want to attempt to reproduce or replicate the results of a previous finding. One is scientific—to establish the extent to which that original finding was robust. I completely agree that there is an excessive focus on novelty and discovery at the expense of ensuring that the foundations are robust and recognising the value of those attempts to check and ensure that robustness.

There is another reason, and it goes back to my point about how we should be thinking of the process rather than the end product, which is that these efforts empirically to estimate what proportion of research findings are, in fact, replicable or reproducible can be thought of as a diagnostic test. If that diagnostic test returns a value that suggests maybe the process is not working as it should, we should reflect on what we can improve about the process and then perhaps run that diagnostic test and look at it again, which is that point of continuous improvement and evaluation that we need to embed in the research process. Most industries will set aside a proportion of their R&D budget on just evaluating the extent to which their processes are in fact delivering what they want them to deliver, and perhaps we do not do enough of that in science and academia at the moment.

The main problem with peer review is that it is done by a very small number of people. Historically, it has been the only quality control check on our process, but there is now scope for far more of that checking through the transparency that we have been talking about that allows data and code and all those intermediate research artefacts to be available for scrutiny.

I do not think there is any single factor that we can identify that is the thing that is the source of problems. It is more that we need to recognise the systems nature of research and all the different contributors to the process and evaluate the extent to which any of those could be improved, even perhaps incrementally. If we can make several incremental changes, we are going to be able to shift the distribution of quality of the whole that we produce much more than by focusing on any one individual thing.

Q40 **Chair:** Professor Bishop spoke of having peer review towards the beginning of the work rather than at the end for publication. Does that not already happen through research grants? There is peer review of

research grants.

Professor Bishop: It is often argued that that is what happens, but it does not really. There would be a real case to be made for looking more carefully at how our funding mechanisms work. I have sat on a grant funding board for the Wellcome Trust. It is very competitive. It is more like a sudden death thing. You have to get your proposal through, and if anybody does not like anything about it, it will tend to be shot down because there are always far more proposals than there are resources. The proposals that go through are the ones that sound most credible, but you do not get an opportunity to improve them.

As a reviewer of those sorts of things, I have often found myself thinking, "Well, they have said they will do A. I wish they would really rather do B, but I dare not say that because it might sink this proposal, and I would really quite like this proposal to go through." It is a rather different sort of mechanism. It could be better if there was a process that could make it more like registered reports. Marcus will tell you they have managed to do that; they are integrating and putting the two together. Who is doing that?

Professor Munafò: It has been done across a few partnerships between journals and funders. It is a model known as registered reports funding partnership. You take the idea that a registered report puts all the peer review in a different place, effectively.

Q41 **Chair:** Just step back a bit. For people watching who are not familiar with the terminology and the points of reference, tell us what a registered report is.

Professor Munafò: A registered report is effectively the first half of what would eventually become a manuscript, where you describe the research question and the methodology that you will use to answer that research question. That is what is reviewed before any data has been collected. What you want to know is whether the question is important and whether the methodology is robust enough to answer that question in a way that means that the answer that you get will be informative irrespective of what that answer is. If the answer is, "No, this thing does not happen as we expected it to," that is worth knowing. We will know that robustly because of the way in which the experiment has been designed, for example. It ensures that the results of the experiment will be useful. It allows for peer review to happen before data has been collected, which then ensures that real changes can be made as opposed to simply putting sticking plasters on an experiment that has already happened.

If we are going to fly an aeroplane, we do our pre-flight checks before we take off, not when we are about to land. That is the basic model. It aligns with what funders do because funders effectively decide whether to award funding, and those that sit on the panels, and so on, on the basis of whether the question is important and the methodology is robust. It

makes sense to align those two processes given that both occur before any data collection has happened.

There are a number of these partnerships between funders and journals that take that publishing format and link it to the grant review process. For example, we are running a pilot between Cancer Research UK and a number of journals in the biomedical sciences, where, effectively, an applicant for funding goes to the funder with a proposal that is the research question and the methodology. If they elect into this pilot that we are running, if they are successful, they are effectively handed on to a journal in a kind of relay race so that the journal can pick that up and review the protocol, which is perhaps the more developed and focused version of their grant that they have already submitted. There are efficiencies there. You have already written your grant; turning it into a registered report is a relatively small step, and you are buoyed by the enthusiasm of having just received some funding.

As well as harmonising that process and making it more efficient, it also allows those reviewers who may have reviewed the grant and thought, "I really like this, and I want to see this funded, but I know that if I say anything critical it is not going to get funded," to come back in at the journal stage and say, "I loved your work. I really wanted to see it funded, but I thought if you did this differently it would be even better." They can make those critical comments then, safe in the knowledge that the funding decision has been made and contribute to the improvement of the study that then eventually runs.

Then what happens is that the applicants have funding to do their work and in-principle acceptance of what will eventually come out of that research process. They can then proceed in the knowledge that there is no incentive for them to overly enthusiastically interrogate their data—get a small p value, whatever it might be—because they have a guaranteed publication in the bag on the basis of that first stage registered report process.

Q42 **Chair:** I see. In your experience and estimation, how much of an appetite is there on the part of funders and journals to collaborate in that way?

Professor Munafò: There is a lot of interest. One of the challenges is that bringing together very different systems—journal peer review and funder peer review, for example—is not trivial. One of the things that we have been doing in our pilot is trying to keep things as similar as possible to how they already run, but there are still issues of how you deal with the confidentiality of grant applications and how you ensure that applicants are aware of what they are opting in to when they are moved across different structures, for example.

In principle, there is a lot of interest, and this goes back to the positive change that has happened over the last five to 10 years, which is that more and more elements of the research ecosystem that I described are

recognising that there is work to be done to improve the quality of what we produce.

The way to do that is to think about innovative ways of updating our working practices, working collaboratively, and joining up different efforts. This is one of those innovations that I think has potential.

We need to evaluate whether these things work as intended. That is important. There are lots of good ideas out there about how we might do things differently, but we need to evidence that ultimately because the irony would be if we started to put in place processes to improve things that themselves turned out not to be replicable.

Q43 **Chair:** That would be an irony.

Let us work on the hypothesis that there was value in this and that it was demonstrated. How would it need to be brought about? Could it happen spontaneously? Could it happen through the networks of influence in academia, or does it require a more muscular intervention on the part of UKRI for public funding, or indeed the Government or Parliament?

Professor Munafò: There is a huge amount of grassroots energy, as we have already said, and that is the real engine of much of the change that is happening. But there is a danger that there will be duplication of effort and perhaps gaps as well, so there needs to be an element of co-ordination as well. For example, the UK Reproducibility Network that we have set up is essentially an academic collaboration that is intended to foster that kind of co-ordination—researchers, institutions, funders, publishers and so on—but we have no real authority in that role; it is just an element of that research ecosystem deciding to work collaboratively.

There is space for some oversight to ensure that that co-ordination is happening and to ensure that that collaborative approach is fostered and, to an extent perhaps, mandated by funders such as UKRI. This is where I see a potential role for the Committee on Research Integrity—that oversight, soft influence role in shaping and co-ordinating that activity, looking for areas of overlap that can be brought together, and looking for potential gaps where new activity is needed.

Q44 **Chair:** Professor Bishop, what do you think are the prospects for bringing about such a change to the system or change to the way things are done?

Professor Bishop: I am broadly optimistic. The difficulty that a lot of people feel is one of time. There is an idea that if you want to work reproducibly and you want to make things open—it comes back to what Professor Ferguson said—everything takes more time; you have to do things carefully. That is the major source of resistance from the scientists from themselves. They feel that they may blight their own careers. It is a bizarre situation where you might blight your career by doing things properly because you are supposed to be doing things quickly and

producing a lot of work. Doing something about these incentive structures is also important.

Academic institutions will have a big role. I know they are interested because I think they regard it as in their interest to change how they function. In so far as they are saying, "If you want to get promotion or if you want to get hired, you have to have a lot of papers in high-impact journals, and you have to have a lot of grant income," that is really rather corrosive to the system because what you really want them to be saying is, "The people we want to support are the people who are doing the careful, in-depth work."

As I think I said in my submission, every now and again an old Nobel laureate pops up and says, "I would never have survived in the current system," because when you look back at some of these guys—and they were almost all guys—what they did was to work for 10 years and perhaps have two publications because they were slaving away really sorting something out in detail. You just would not be viable if you did that sort of thing these days. I am not saying we should go back to that necessarily. A model that involves much more collaboration and different groups working together to solve problems is one way forward. Traditionally, in recent years that has not been valued.

What has been valued is that you show that you are the person who had the original groundbreaking idea rather than that you are a good citizen who works well with other people as part of a team. Most of the really complicated problems that we are trying to solve these days really require teams with different expertise working synergistically. We have to find a way to recognise that.

Professor Ferguson said it all. He was talking about the need for software engineers. I would also add applied statisticians who will work with groups. These are at the moment not very viable career paths, but we need people in those roles if we are going to make some of this change happen.

Q45 **Chair:** Against all that, speed is important, is it not?

Professor Bishop: Yes.

Q46 **Chair:** We have seen that through the pandemic. It is a line that our colleague, Aaron Bell, has been pursuing in vaccine development.

Professor Bishop: Yes.

Q47 **Chair:** Fantastic and remarkable though the speed has been, it might be possible that we could have been even speedier and saved even more lives. What you are describing might be a Rolls-Royce system of making sure there is proof against irreproducibility—that is a hard enough word without adding a negative to it. Are we not in danger of prizing perfection against the impacts—in some cases, life-saving impacts—that come from pace?

Professor Bishop: There is always going to be a trade-off. I was very interested reading Sarah Gilbert's book on Vaxxers. She said that the reason they did things much faster than they used to do them was not because they were cutting corners; it was because they were able to do a whole load of steps simultaneously that they normally would have done sequentially, pausing to seek funding in between each step. She was at pains to say that they had done things equally carefully.

I am not an expert on Covid, and I try not to say too much about it, but my impression is that it has illustrated, if anything, the dangers of going too fast and cutting corners. There has been an awful lot of really bad Covid papers. It has revealed the worst and best of what science can do.

The worst has been highly destructive in making people think that various drugs and so on might be effective, which have turned out not to be, but, meanwhile, the genie is out of the bottle, and people are rushing around and trying to apply them. You need to be very careful.

What we certainly do not want to do—and it is a real risk—is increase bureaucracy. I know you have another Select Committee on bureaucracy. That is hard if you are going to say you want to do things more slowly, more carefully and audited more. You are trying to balance these very difficult things.

I certainly do not think there is much value in putting out poor-quality research because of this business of people building on things, particularly if they look really exciting and sexy. You continually meet PhD students who spent two or three years trying to build on something only to find that they cannot reproduce or replicate the main finding. They initially think it is their fault. They can spend a long time wasting their time and energy only to find that other people say, "Oh no, we couldn't do that either." It is really important to recognise that, if you allow bad-quality stuff to get out, it is not just that there is one bad paper out there, but it can seed a whole line of research for years that is wasting resources. It is a real waste of time and money.

Q48 **Mark Logan:** Professor Bishop, on the point about working in silos, do you think that is perpetuated by our undergraduate system? I am thinking of the United States, where it is very interdisciplinary. From ages 18 to 21 they do all sorts of different modules as part of the course, whereas when I was doing my undergraduate degree of 18 modules, 17 were in that specific subject area of law that I was studying. Does that feed into your point?

Professor Bishop: I had not thought of it in that way. The trouble is that there is so much we all need to know it is very hard to branch out even beyond one own's narrow area. Scientists tend to divide into those who get highly specialised and very good at one thing and those who attempt to cross disciplines. I have been one of the latter. I try to dabble in various areas that involve the subject on which I am working, but it is

very hard. You can easily become unstuck because there is too much to know.

We have got to the point where many scientists feel they have to know so much stuff and they cannot keep up with all of it. We have to accept that we need to specialise in our own areas, but then get people together across the different disciplines and build those bridges.

I genuinely do not know whether you could start doing that at a university level. To try to persuade people to cross disciplines is an interesting idea, but I do know that most funders like the idea of interdisciplinary or cross-disciplinary research, but, at the same time, when people submit grant applications across disciplines usually they do not go through because neither of the disciplines really likes them. If you want that kind of cross-disciplinary engagement you have to work quite hard on it; it does not seem to happen naturally.

Professor Munafò: It is a good point and is something we have thought about. That is where all researchers, academics and undergraduates start. There is a place for thinking about foundational skills, not so much across disciplinary boundaries, although what I am about to propose might facilitate more of that cross-talk across interdisciplinary boundaries.

To step back, we talked about data sharing, code sharing and open research practices. We have excellent digital infrastructure for that, particularly in the UK. Many institutions have data repositories and so on, but you also need the soft infrastructure—training and skills—to be able to do things in the right way before you make a deposit. That needs to be co-ordinated so that the way it is done in one institution is the same as it is done in another, because researchers are mobile and move across institutions. Inefficiencies are created by things being done in different ways across groups, institutions and even across countries, if you want to zoom right out.

Perhaps the place to start with those foundational data skills, if you like, is at undergraduate level. You bring together researchers from quantitative biomedical disciplines, regardless of the specific degree they are doing, to understand the basics of how to put together a spreadsheet and create a good data dictionary and the metadata that goes with your deposit, and so on. These are skills that will be relevant when people go outside academia, as most undergraduates will, but they also mean that all those who remain within academia have been trained in broadly interoperable ways that provide that foundation, which then makes it easier to adopt the open research practices as they progress through their academic careers. I certainly think there is a place for incorporating undergraduate training into this bigger picture.

Q49 **Mark Logan:** I am going to shoehorn two questions into one given time constraints. Professor Bishop, you touched on this in your earlier response to Becky. To what extent is irreproducibility a problem for all

research rather than specific academic disciplines?

The second part of my question as a budding social scientist is that we always get hung up in trying to have causal mechanisms. If something does not have a causal mechanism, is it useless? Should you not pursue the research? Those are two parts to that question.

Professor Bishop: Given that your briefing specifically asked about this, I was looking through it and thinking what areas I know about. I was thinking initially that there are fewer issues around physics and chemistry, the harder sciences. Then I found that was not necessarily the case. There was a recent paper in *Nature* on real problems in quantum computing where researchers thought they had discovered something called the Majorana particle. Then somebody else said they had not really discovered it; they just thought it was there. People were seeing things in that experiment things that others could not replicate. It seems that the problem is there.

Earlier the Chair asked about AI, machine learning and so on, and what that can tell us. As far as I can see, that area is fraught with problems and things do not replicate. Everybody has their own little machine-learning system that they put out there. A paper earlier this year in, I think, in the machine-learning *Nature* journal, said it was looking at machine-learning algorithms that were trying to predict who would recover from Covid, or who would get particular symptoms from Covid. None of them replicated.

Machine learning is good for generating hypotheses, but you need another dataset to test the hypotheses. It comes down to the idea of distinguishing between exploration and homing in and maybe firming up a proper theory on the basis of what you have done. It is certainly not a panacea. It is widely used and is the sort of thing that people think they need to do currently, but I have not been impressed from what I have seen of people trying to use it to diagnose the disorders I work on, for example, on the basis of brain scans. It has gone nowhere.

Clinical medicine is interesting. Clinical medicine and clinical trials were first off the block in noticing problems, particularly problems in publication bias and people shifting the goalposts in how they did their analyses. We had some excellent people in the field laying down how things should be done. We now know how to do good clinical trials, what analyses should be done and so on – and that you still have people saying they should be registered. I think that has improved since the last time I was before this Select Committee, but it still seems to be the case that the people working in clinical trials are pretty unhappy that they do not all get registered, or when they do they do not get published. There are still a lot of unpublished clinicals trials and there appear to be no sanctions on that.

As for leading the way and giving us a lot of ideas about how to do science, that field has been important, perhaps because the stakes are

high. I have already mentioned psychology. It has been through a terrible crisis and I think it is coming out the other end, which is good. Biomedicine has a lot of problems similar to the ones in psychology, but less well recognised.

Economics is interesting. Economics looked at itself. It is so different because, although it has experimental economics, a lot of what it does is analysing existing databases, so there really is an issue of reproducibility in its formal sense rather than replication. There have been some analyses in that field showing that a lot of the published studies in economics could not be reproduced. People would take the same dataset and not get the same results.

At that point—I think it was about five years ago—economists woke up and their top journals then started to demand that code and data be made available. That was their strategy for dealing with it. My guess is that things will have got better, although I do not know whether anybody has done a formal analysis.

Genetics had a big problem around the turn of this century. Initially, they had not understood data dredging and so they were coming with all these associations between this genetic variant and that condition. They then realised that they had to adopt completely different statistical approaches to deal with it, but also typically to require that for anything you try to publish in that field you must also have a replication.

Different fields have come up against this and invented different ways of tackling it. That is one good reason why it is quite nice to have something like UKRN, which tries to pull together different disciplines to see what problems they have uncovered and what solutions they have found.

There are fields like cancer biology or cell biology where some of the problems may be due to just not having the reagents you think you have and you need to test whether, when you order a reagent or cell line from another lab, it is what it says it is. There have been notable instances of these things not being checked and being wrong. In that case it is more to do with lab protocols being properly specified and made clear.

Each of these areas has started to address the problems. The most severe problems are not necessarily the same across all areas, but it is helpful to compare and contrast, not that we can impose a solution from one area to another but rather that we can learn from each other.

Q50 **Graham Stringer:** This has been a really interesting session. There are hundreds of questions I would like to ask, but I will restrict myself to two.

When Professor Slingo, I think it was, came before the Committee she said the Met Office was buying a supercomputer. She said that one of the problems was that if it put the same data and codes into a new supercomputer it did not necessarily get the same results. Is that a problem with which you are familiar—that the results depend on which

machine you use?

Professor Bishop: Fortunately, I do not work with supercomputers and I am very relieved that I do not. We have had issues in things like brain imaging. You do a study on one brain-imaging system and you find somebody else tries to do the same one with their own system and gets a slightly different result, or sometimes a very different result, because of the machine settings. These are things one has to be concerned about and get to the bottom of. You need scientists who get these different results to put their heads together and work out what the source of that difference is, but when you are dealing with highly complex systems this is very much the sort of thing that can happen. A tweak of one little setting somewhere may have a knock-on effect way down the system, but it is not my area of expertise. That is all I would say.

Q51 **Graham Stringer:** I go back to the question Aaron asked of Professor Ferguson. We have been talking about fallibility in the systems. How big a problem do you think out-and-out fraud is in the system? The whole of science has never really owned up to its failures to deal with the Wakefield MMR paper. It took a long time for the *Lancet* to deal with this. It was a journalist who sorted it out. Do you believe that cheating and fraud is much of a problem, or is it a once-in-a-decade problem?

Professor Bishop: The standard answer most people will give you is that it is not a big problem. I am less complacent. We do not necessarily pick it up if people are clever. The ones we pick up are the ones who are stupid. There was a wonderful case where a guy who had invented his data in an Excel spreadsheet left in the formula he had used to generate the data. This is not an intelligent way to behave.

I think this will always be related to incentives. You will see the biggest evidence of fraud in places where there is very great reward. There was a time—this may have stopped now; Marcus may know—when for certain Chinese institutions, if you published in *Nature, Cell* or *Science*, you got the equivalent of a year's salary as a bonus. That is a strong incentive for fraud.

I have spoken to people in other countries where sometimes in some places whether or not you are rehired depends on whether you get a paper in one of those top journals. These are very high stakes. If we want to get rid of the fraudsters, we have to get rid of the high reward for the proxy indicators of quality and start rewarding genuine quality on metrics that cannot readily be gamed.

I share your concerns. Fraud is probably widespread, particularly in the context where it is very high stakes for fraudsters. There is a famous case involving a Dutch fraudster, Diederik Stapel. He wrote his biography about the process. Sometimes people get annoyed about this and say he should not benefit from it. I do not think it is even a published book; it is a document available online, but it gives a fascinating insight into the mental set he went through. He was quite a successful psychologist in the

Netherlands, but he was getting rather boring results. He found that if he got an interesting result everybody was all over him and it was all great, and then he started tweaking it a bit and it got worse and worse. In a way, he had entered into a Faustian pact and he was dragged down. In the end, you felt quite sorry for him, which I am sure he intended, but it was probably not an appropriate reaction.

This is why I worry. If early career researchers get pressure from their bosses to do little things, even a bit of p-hacking, which does not seem so serious, it is quite a corrosive influence. These are people who typically start out idealistic. You do not go into a science career to make a lot of money usually; you go into it because you love science—it is fun and you find it intellectually stimulating. To see people being gradually turned around and feeling they cannot have a successful career unless they put out incredibly exciting data is depressing.

Professor Munafò: I completely agree with that. The reason for all the grassroots energy we have talked about is that there is an increasing awareness among early career researchers. No one goes into academia thinking they will game the system and become moderately wealthy in 20 years' time. That is not the motivation, but people get bent out of shape by the incentive structures.

We need to distinguish between researcher integrity and research integrity in the integrity of the research and the research process. Those are distinct things. Many of the things we have been talking about—research transparency and incentives—will improve the integrity of the research process and, at the same time, make it much harder for people to cheat and game the system. Therefore, we can focus on one for much more general benefit while addressing the other. Effectively, I take a public health stance on this. Rather than focusing on the rare cases that might have individually a disproportionate impact but are rare, we should focus on the more widespread issues where, if we can address them in a more mundane way through things like research transparency, we can shift the distribution. It is a bit like everybody drinking a little less alcohol and the population as a whole being a lot healthier. That kind of approach will gain more traction and move us further when it comes to improving the quality of the work we do. What we need to do is move away from a system and culture that focuses very much on trusting individuals and individual researchers to a process that is inherently more trustworthy because it is transparent, for example.

Q52 **Aaron Bell:** Professor Bishop, you mentioned clinical trials and publication bias. This Committee's predecessor reported on that. In this inquiry we have received evidence from Cochrane, Transparency International, Global Health Programs and TranspariMED, which is No. 24 on the website. They recommend we should put in place some sanctions with teeth for sponsors that fail to make trial results public; that we set a starting date for that; and that we conduct annual follow-ups until that improves. First, would you recommend that for clinical trials? Secondly,

could we extend that to other disciplines? You put in place sanctions for people who do not report results they have been conducting.

Professor Bishop: The answer is probably yes. I am always a little guarded about jumping in and saying that is a great idea, because there are so many areas where there are unintended consequences. It might be worth trying that and seeing what the impact is. Marcus is probably better able to answer that.

Q53 **Aaron Bell:** What happens if you register a report and do not publish the second half of it?

Professor Munafò: The point is that it is transparent. You can see that, so you know something has happened and you can go and ask them. That transparency allows you to go in and check. In the context of clinical trials and data sharing, the reason we have a concern is that we know data are not being shared. We know that because studies are pre-registered. One of the reasons in part that the vaccine programme has been so successful and rapid is that those things that we have been talking about are already embedded in that particular area of research in multi-centre collaboration, pre-registration, data sharing, etc.

As a psychologist interested in behaviour change I am always wary of sanctions as a solution, but we can look at whether the existing incentives within the system can be modified to promote the behaviours that we think would make a difference. For example, the future research assessment programme is ongoing to look at what the next research excellence framework might look like. The previous one focused very much on outputs and much less on environment. One thing that could make a big difference in the context of what you are describing is much greater focus on environment and process and much less on output. For example, what proportion of papers generated by an institution has a credible data-sharing statement?

Q54 **Chair:** There is one final question to Professor Bishop, perhaps a provocative one. Going back to the discussion we had about machine learning, that seems to me to be likely to advance in all areas of technology in our lives. As you rightly point out, the essence of it is irreproducible in the sense that the machine sets off and makes connections without hypothesis; it informs itself, makes other connections and comes out with results. In these discussions, are we not a bit like King Canute? We are addressing some concerns that are about to be swept aside by the power of the insight that comes from the analysis of big data by machine learning that will be impervious to the discussions we are having.

Professor Bishop: I am not an expert on machine learning. All I do know is that, where it has been tried in areas that I know about, every machine- learning study seems to get a different result and they do not generalise. I know it has been highly effective in things like visual recognition from scenes and playing chess. Some studies that involve

that approach have managed to be highly effective, but often it is just applied as a kind of hope and a prayer that a bit of blind analysis will give us much more insight than anything more theoretically constructed or from a hypothesis.

I would say that so far it has not lived up to that promise. You will find random patterns as well as real ones and there should be methods for dealing with that, but at the moment people are perhaps not throwing out the random patterns quite enough and making a big deal about them.

Chair: We are very grateful to both of you for coming to give evidence to us. Thank you very much indeed.

Examination of witnesses

Witnesses: Dr Oransky and Dr Clayton.

Q55 **Chair:** Our final panel of witnesses joins us from the east coast of the United States where it will not have escaped our attention that it is 6.15 am. We are especially grateful to our next two witnesses for joining us this morning.

I am very pleased to welcome Dr Janine Austin Clayton, associate director for research on women's health at the National Institute for Health in the United States. Dr Clayton has been instrumental in exposing questions of gender bias in research in the medical field in the US.

I am also pleased to welcome Dr Ivan Oransky, who is editor-in-chief of the journal *Spectrum* and teaches medical journalism at New York University's Carter Journalism Institute. He is the founder of and responsible for the blog Retraction Watch, which scrutinises the retraction of medical papers.

Thank you very much indeed to both of you for appearing before the Committee today. We are kicking off our inquiry into reproducibility in research integrity. Starting with Dr Clayton, perhaps you might give us your views on how much of a crisis or problem is being faced. Some people say it is a crisis; some say it is not. What is your opinion, Dr Clayton?

Dr Clayton: There is certainly an issue when there is a failure to reproduce very large studies that have been repeatedly reported over time in a lot of different settings in different scientific disciplines. In the area of clinical trials the failure to reproduce these large studies has created a lot of attention. There is also a concern around eroding public trust in science's ability to solve society's problems. I do think it is a problem. It is being addressed in forums like this and in others in other countries in a variety of ways. We are making some progress, but it is indeed a problem.

Q56 **Chair:** Dr Oransky, from your particular point of view and the scrutiny you give to the retraction of medical papers, what do you see as the trends here?

Dr Oransky: It is important to understand the iceberg effect, if you will. We heard from the previous panel, to which I had the pleasure of listening, that one important distinction is between fraud, which by definition is intentional, and the larger reproducibility issues.

I do not like the word "crisis" for a number of reasons. I think it is very polarising. I am by no means in denial about the size and scale of the problem, but the word "crisis" implies that somehow it is a new problem. One of the things we have seen with retractions, which I will speak about briefly in a moment, is that what is good about all of this—that is probably a counterintuitive thing to say—is that far more people are looking for these problems, and essentially we are seeing a screening effect.

To use a clinical analogy, if you never looked for, say, a particular cancer or other diagnosis you were trying to note and treat and did not have any screening test for that, it would appear as though there were very few problems or very few cases of that particular cancer. There are issues with cancer screening, which I will not go into because it goes far beyond the scope of this, but, looking at it as if you add a screening test to the standard regimen or the guidelines, you will find more cases.

Are there actually more cases, or are we simply picking up a lot of cases that we were not picking up before? To me, that is the thread that connects a lot of what we are seeing in terms of reproducibility, which, as Professor Bishop very accurately described, is happening, as best we can tell, in every field, and retractions. The number of retractions has grown a great deal.

To step back for a moment, retraction means that a journal, an author or an institution is saying, "This paper is no longer reliable." It can be for any number of reasons; it may arise from fraud or misconduct, but that is what it means. Earlier this morning I looked up some figures to remind myself. There were about 40 retractions in the year 2000. In 2020, according to our database, which we and many others consider to be the most comprehensive for retractions, there were about 2,700 retractions. Even though the number of papers published has certainly grown in that time, it is not the same ratio: 40 to 2,700. In case it is of interest, in the UK there were three in 2000, 75 in 2010 and 96 last year according to our records, which we think are comprehensive but may be somewhat incomplete.

We are seeing many people looking for these issues, which we think is critical and, frankly, terrific. We call them "the sleuths". Some people do not like that term, but these are people who are finding things, whether they are reproducibility problems, out-and-out fraud problems and everything in between. It is important to note that we are not picking up all the problems yet, and that may also be a subject of discussion.

Q57 **Chair:** Both of you obviously have experience of the US system but a perspective on other countries as well. Do you see any particular country

that is taking action in advance of others and seems to offer benefits to, in our case, the UK system?

Dr Oransky: I think it is worth looking at every country and perhaps certain examples. In the US, as Dr Clayton knows, there are long-standing—by which I mean decades—agencies and oversight bodies. I speak here at a high level, and if there are any questions we can dive in. They have been in place since the late 1980s and early 1990s. One is in health and human services, of which NIH is a part. That is the Office of Research Integrity. The National Science Foundation has a similar organisation; it is an office of the inspector general. There are differences, which I am happy to discuss but are not particularly relevant.

They are often pointed to and cited as some of the earliest and most effective agencies and oversight bodies with the most teeth. A lot of countries are either directly or indirectly starting to emulate that, as I know, because I also provided oral evidence in 2017 to this Committee. I do not want to overstate the case, but, if I understand correctly, a report led to some steps that were in that direction but not necessarily in terms of investigative bodies.

This is a question that many countries are asking. The Scandinavian countries are taking some interesting steps and have been for some time. One of the things to note, however, is that for all these steps their antecedents tend to be a bad case or a number of them. That is natural; it is human nature.

On the other hand, to get back to what I heard from the previous panel, a forward-thinking approach to get ahead of the problem—by the way, I do not know whether we are ahead of it any more, but at least we are aware of it now—is to me a good way to think about it.

Dr Clayton: I would agree it is important to look at different countries' efforts. In my area, we have looked at how different countries have implemented various strategies to address the role of sex and gender in rigour and transparency as well as reproducibility. We have looked at different countries and collaborations with a variety of groups.

Different countries have taken different approaches to the problem in terms of which step they address first. They are on a different part of the journey. I think there are lessons learned, but the systems are different.

I heard from the previous panel about the issue of perverse incentives and the pressure on scientists to publish. That part of the structure is quite common across all countries for academics and scientists. Looking at different ways in which different countries may have addressed this is germane to the issue and can be very useful.

As for the disincentives, incentives and the perverse pressure to publish and structures, there are ways to do some comparative analyses to see what might work best in different settings.

I agree with Dr Oransky about the importance of being proactive and even studying interventions in this space so that we can generate evidence and share it internationally on what has worked where, for whom and in what setting. I hope we will get a chance to talk a little more about some of the specifics, but that is my response to the question about comparing countries.

Q58 **Aaron Bell:** Dr Clayton, following what you have just said, do you think there is a single biggest cause of irreproducibility, and, if so, what is it? We have just talked about incentives generally. I do not know whether you heard Professor Bishop's answer in the previous session, but she basically described it as undergraduates wanting to produce good results, to advance and so on and something that builds up through the system. Is that how you see it, or is there a different single biggest cause?

Dr Clayton: I am not sure I can say there is a single cause, but I do think that that is an incredibly intense aspect and factor that scientists today are dealing with, not just at student level but at the academic level. The pressure to publish, get the next grant and be promoted and tenured are all important factors that are drivers. It is an important driver of what we are seeing.

We may be unaware of other factors, and unless and until we examine that more carefully—we are doing that in efforts like this and others who are examining what is beneath that—we may not know some of the other drivers. I think it is an important one, but I cannot say it is the most important one. There was a discussion previously about the role of fraud. It is very hard to know how much of that is contributing. Most people believe that it is a relatively small aspect of what is causing the challenges to reproducibility. As we dive deeper and examine more of these papers, we will learn more over time.

Q59 **Aaron Bell:** On a related point, do you think that the peer review system works in the interests of reproducibility? Is peer review sufficiently robust, or is there too much very light touch reviewing, because in most cases it is a voluntary job and people do not have time to go into the details of the statistics or the code?

Dr Clayton: In the United States I think the peer review system is incredibly robust. Individuals take that service very seriously. We have expanded the requirements for peer review to address the scientific premise and the rigour of the application, and that is where sex as a biological variable in gender as a social construct comes into play and how well the application comports with the scientific method to ensure unbiased and well-controlled experimental designs, analysis, methodology, interpretations and plans for reporting a result.

Peer reviewers have to evaluate a lot of things, not just this, and there are many applications that are reviewed in each round, so there is time pressure for them to review a lot of things and discuss them. Peer review continues to be a solid component of science policing itself, but in and of

itself it cannot address this issue alone, because it is not just about peer review of the application or peer review of a manuscript. We have the important issue of journal editors and publishers who are critical gatekeepers of this process in science to get the information out to the public. Unless and until that is routinely considered—from my perspective, looking at sex as a biological variable in gender and ensuring that the information somebody would need to try to reproduce a study is included in the publication, either in the ancillary materials, the methods or the supplemental materials—we will not even be able to reproduce every study because the information is not provided. The sex of the animals or subjects and the breakdown by treatment and control groups for randomised controlled trials is information that is not always provided, so we have an issue with rigour and transparency that affects reproducibility.

Q60 **Aaron Bell:** Dr Oransky, you have given us the statistics for the number of retractions. How many of those could have been avoided by better peer review at the time, and how many arose from things that came to light afterwards?

Dr Oransky: It is a great question. The simple answer is that 100% of those cases came to light afterwards, but that does not necessarily tell us how many could have been prevented. Dr Clayton, you will correct me if I am wrong, but I believe you were referring obviously to the grant peer review process. You discussed journal peer review towards the end of your response. I want to be clear that we tend to focus much more, as the name Retraction Watch implies, on the peer review process that is happening in journals or the peer review process that, frankly, is not happening in some journals.

One of the issues is that we have come to believe or have been told or instructed in some way—it is part of the ether—that peer review is not magical but, as we would have said in the States in the old days, is the *Good Housekeeping* seal of approval. I give due respect to the people who work incredibly hard at peer review. By the way, I am one of them. Along with my colleagues at Retraction Watch, I peer review a few papers a month. I know that is not all that many compared with what most scientists do, but we do take part in this process and think that peer review is important.

However, we think that the false binary of it being peer reviewed or not peer reviewed is part of what is creating these issues. There is peer review and there is peer review. There are journals that do very rigorous peer review, and sometimes they do not; there are journals that hardly ever do any rigorous peer reviews, and sometimes they do; and there is everything in between.

To give a personal anecdote, I trained as a medical doctor in the States, but I do not practise medicine. I am not an immunologist; I am not a virologist. I could go down the list, but my point is I am not a Covid-19 expert. Yet, in February and March of this year I was asked to peer

review five different papers about Covid-19. These were journals of a leading publisher; it is a very large one we have all heard of.

I would ask you how that came to happen. Going back to the previous panel, it more than likely came about because of algorithms. I am in the database as someone who has published a letter four paragraphs long about Covid-19 retractions, so I come up as an expert on Covid-19. That is absurd.

To be clear, particularly given this setting, I did not peer review those papers; I declined the invitations. The problem is that not everyone does. It is an honour to peer review; it is part of your tenure, promotion and committee work—in other words, you are supposed to consider best service to the community as being very important. It is really important, but the problem is that too much does slip through the cracks, even among things that are peer reviewed. When it is not peer reviewed, that can be even worse.

If I may be positive for a moment, one of the things we have seen during the pandemic is how quickly a lot of these issues are being corrected. It does not happen always and, frankly, it still does not happen quickly enough in many cases, but because of the nature of preprints, which I imagine we will discuss at some point if that is of interest, these are labelled as not peer reviewed, yet they are often corrected much more quickly or even withdrawn or retracted. "Withdrawn" is typically the term that is used. That tends to happen quite quickly.

To me, the important issue is not whether it has or has not been peer reviewed—we should know that—but what is the nature of the peer review? Some journals are taking various steps that basically ensure trust in that process—for example, posting peer reviews regardless of whether they use people's names. There are a lot of issues with using someone's name. That can be a problem for someone's career if they are critical and someone else does not like what they have said.

On the other hand, as regards anonymously posting peer reviews, it is early days. To echo what we have heard from a lot of folks today, we want to see whether these things work; we want to be empirical and evidence based, but that is certainly one thing that is worth trying so we can increase and ensure there is trust in the process and make it transparent.

Q61 **Aaron Bell:** You have anticipated my line of questioning on Covid and preprints. Do you think it is possible at this stage to conclude whether the research that has been conducted over the course of Covid—I appreciate you are not an expert on Covid—has been more or less reproducible than is normally the case? Has there been more fraud than normal because of the high stakes involved?

Dr Oransky: The simple answer—and, I believe, the correct one at this time—is that we just do not know. To me, that is evidence based. We simply do not know.

What do we know? We know that as of yesterday—maybe there are more this morning that we have not yet seen, because it is early here—194 Covid papers had been retracted. That seems like a large number. Even last year, there were 2,700. These numbers seem large if you do not follow them; they may seem small if you do.

As a percentage of papers published—this includes preprints—it is, if anything, a little smaller than the overall rate of retraction over time. The caveat—and the reason we do not know yet—is that retractions tend to take on average about three years. It is not a smooth curve and is probably bimodal; in other words, there are some that happen quite quickly and others take decades. Because they happen on average in three years it means I would want to wait at least three years from when they were published. Of course, we are not there yet; we are not three years into the pandemic, even though for many of us it feels much longer. You would really need to look at that.

This intense scrutiny, which I think is really good for the process and science, has led to things happening much more quickly. We cannot say with any certainty whether the science is more or less rigorous. The fact is that a lot more of it has been done in an open way and the scrutiny has been out there much more publicly and frequently. My prediction is that that augurs well when the dust settles for looking at the scientific record, but I should never make predictions and I want to be careful there.

Q62 **Aaron Bell:** You refer to preprints. There have been a lot from Covid because of the timescales we are dealing with and the pace at which we need to operate. Do you think that overall preprints contribute to a more reproducible academic knowledge base? Would you encourage more preprints in future in different fields?

Dr Oransky: I would. It is important to keep in mind that preprints are not new. Thirty years ago our archive was created. Paul Ginsparg has just won a major new award at the Einstein Foundation in Berlin for his work on that, which continues to be seminal. On the other hand, BioArXiv and MetaArXiv are much newer. BioArXiv dates from about 2015; MetaArXiv, which tends to publish probably the most relevant material for Covid-19, given it is publishing mostly clinical trials and things that are clinically relevant, did not launch until months before the pandemic in 2019, so it is to be expected that these have picked up for obvious reasons, as you point out.

Overall, I think they are an excellent thing for science. There are concerns about whether they are labelled properly; there are concerns about how they are used downstream; there are concerns about bad actors very occasionally trying to publish things that they know are really

problematic—sometimes they do not know they are problematic—but there are also some good examples.

I give you one example. To be clear, others may not believe this is a good example. The Ottawa Heart Institute posted a preprint; it claimed in a paper that there was a link between one of the Covid-19 vaccines and myocardial heart inflammation, which is a very serious condition. This took off in many circles that were trying to raise doubts about vaccines and it was very problematic in that sense. Within a few weeks, having been notified of the problems, the authors withdrew it instantaneously essentially. That was again a story of rapid correction. I should have mentioned it turned out that they had made a massive calculation error that became very clear as soon as people looked at the data.

That was not peer reviewed. Would it have been caught in peer review? I think that is the assumption, but it was caught during what is called post-publication peer review. In this case, it was pre-publication peer review or public peer review. It underscores the fact that preprints can allow that to happen. Some journals are picking up on that and starting to work with it, but journals typically have not encouraged it, to say the least.

Q63 **Aaron Bell:** Dr Clayton, from the NIH perspective what have your observations been about research during Covid-19, and the reproducibility and integrity of it? What are your views on preprints?

Dr Clayton: The Covid-19 pandemic has certainly presented an incredible and important opportunity to invest in rigorous responses and responsible research that will help us to understand clearly how to prevent and treat disease in everyone. By that, I mean that it is critical to consider male and female differences and social determinants of health that have been well documented to affect outcome in the United States.

The challenge the world is facing has presented unprecedented—I hate to use that word over and over with Covid— opportunities to increase rigour. Rigour needs to be embedded in the process so we have confidence in the results as well as an understanding that science is iterative. We may learn new information that renders the points made in previous publications moot. The new science builds on previous data. That is how science can move forward.

Preprints are an important way to get information out to the public. This kind of public peer review with preprints can be very useful in challenging notions, methods and information that is presented, but is not a substitute for subsequent publication. They are one way to get information out quickly. We have learned during the pandemic how critical it is to have shared data and information as widely available as possible to advance our understanding of the SARS-CoV-2 infection and the pandemic.

We have created some sex and gender guidelines for Covid-19 researchers in which we highlight why it is so important to look at well-

recognised male and female differences in immune responses. Women tend to have more robust immune responses than men, and we are seeing different patterns in mortality. A greater number of men die than women in some settings, so we are seeing sex differences, making it very clear that we need to study them and embed that in the design of our studies, analysis plan and reporting.

I want to add to my previous comments. There are journals that have adopted the sex and gender equity in research guidelines—SAGER—developed by the European Association of Science Editors' gender policy committee, and other journals have adopted their own policies. We do see improvements in reporting and analysis. Unfortunately, that is not across the board and it differs by discipline, especially in the pre-clinical space—those studies that lead up to the clinical studies. We know it is so important to have rigour at every stage of the research continuum, especially as we are noticing with Covid-19 and the pandemic that the animal studies that lead up to the clinical studies need to be performed as rigorously as possible, meaning they need to include males and females; they need to have masking and randomisation and all of the tenets of a robust scientific method.

Preprints play an important role and can help us between that discovery phase and dissemination once we have publications that are then reproduced. It is an important part of the ecosystem, but what is critical is our ability to review, think quickly and analyse the information in the preprint, which speaks to the ability of scientists to use the skills mentioned by the previous panel. Are we training undergraduates, medical students and scientists in rigorous experimental design and statistics and in how to interpret the results and understand the limitations and caveats in a preprint or in a publication?

Q64 **Mark Logan:** Dr Oransky, given the pressures placed on researchers by the academic system, is it possible routinely to produce reproducible work?

Dr Oransky: It is certainly possible because there are researchers who are producing reproducible work. I think the question is: how much of it, at what cost in resources, and are we prioritising it?

Publish or perish is a real issue. We heard from the previous panel on this. It is a welcome development. Professor Bishop referred to this. In China, what we think of as the most perverse incentives, even though they are the same incentives as elsewhere except that the scale is much larger, have been banned. Professor Bishop referred to a very large bonus being based on what is known as the impact factor. It is not really a ranking, even though people use it that way. How often are papers are cited in a particular journal? They give it an impact factor and multiply it by a certain currency figure and you get a large bonus.

That has been banned as well as what I think of as a more pernicious problem, which is not so dissimilar from what happens in other countries

but again happens on a much greater scale or order of magnitude. Clinical faculties at medical schools in China were expected to publish two papers just in order to keep their jobs, get promoted, etc. That is very unusual. I can tell you as someone who has graduated from medical school that pretty much all the clinical faculties in most parts of the world are expected to teach and train residents and medical students and treat patients. They are not expected to do research, so you are setting things up, if not for fraud, which does happen there and elsewhere, certainly for failure or lack of rigour and reproducibility.

As to how to fix it—we heard some good ideas from the previous panel—start upstream. I do think that sanctions are important in many ways; they are necessary but not sufficient. If we start to look upstream and whether it is the incentives, the UKREF is something to examine. What does that incentivise? That is not dissimilar from what happens in other countries, but every country is different, as we heard from Dr Clayton. We start to think about some of those incentives, but when we look at the pandemic it is not necessarily publish or perish, although that is certainly part of it. People want to be first; they want to have papers in major journals; they want to be cited later on, but there is an honest, almost altruistic reason for wanting to get results out there. We all want this pandemic to be solved, over with, or whatever the word is; we want treatments and vaccines and all of those things. We want to understand it better. That is a positive.

That needs to be balanced against the need for rigour and reproducibility. I do not know whether "solved" is the right word, but if we understand the context better and understand that knowledge is provisional and iterative, which was one of the words used by Dr Clayton, people must understand that so that there is not the false binary that this is true and that is not true. No. This is the evidence for what we think and, therefore, we are pretty convinced that we should act based on that evidence. That is different from saying something is true because it appeared in a journal and a researcher was under pressure to publish it, that being quite literally the coin of the realm. That is something we should be thinking about.

Q65 **Mark Logan:** Dr Clayton, would you deal with the same point?

Dr Clayton: Will you repeat your question?

Q66 **Mark Logan:** Given the pressures placed on researchers by the academic system, is it possible routinely to produce reproducible work?

Dr Clayton: Yes, it is possible and it is being done. With careful attention to providing the detailed information, the methods and all the details and specifications of performing the experiments or conducting the studies, it can be done. It is important that we have policies around data sharing, including ensuring that you share enough data so that someone can replicate your results. NIH has a data-sharing policy and a new, expanded one will be coming out soon. We will be releasing that

information and it will be publicly available. Sharing data is critical and it can be done.

I want to emphasise that science is iterative and our understanding evolves as more evidence is generated and shared and we critically evaluate it, formulate policies, design new experiments and generate new results. It is similar even from a clinical perspective. I do believe we can do that. We are making progress and the pandemic is upping the ante of how important it is to be rigorous about it.

Q67 **Graham Stringer:** Dr Clayton, I think you said at the beginning that the peer review system in the United States was thorough and rigorous, or words to that effect. Does that mean you think it is better than in other countries, particularly the United Kingdom?

Dr Clayton: No. I am just speaking from my experience. That is my experience of the US system. I do not have any experience of the UK system, so I would not be able to compare it.

Q68 **Graham Stringer:** I just wondered whether the peer review system was a universal system or whether it varied around the world.

What proportion of medical research would you estimate is based on biased samples?

Dr Clayton: I wish I could answer that question, but because the information on the sample is not provided it is not reported from a sex perspective. It will not give the numbers of males or females; it will not say which sex was used. Even in clinical publications, fewer than one third of NIH-supported phase 3 large clinical trials eventually have a publication that reports results for women and men separately so that we can see those results distinctly. Therefore, we do have a problem. I cannot even answer that question because of unreported data. I wish we could have that information because it would give us critical insight into what happens, especially in that pre-clinical space, but when investigators do not report the sex of the animals or research subjects we just do not know.

Q69 **Graham Stringer:** In terms of people, is it possible in theory to recruit a completely representative sample?

Dr Clayton: For clinical studies, we aim to address the scientific questions under study in a particular clinical protocol, understanding what is representative for that disease and the prevalence differences that we see by sex. There are diseases much more common in women than in men, or vice versa; there are racial and ethnic-minority differences; and there are social determinants of health. There are many things that affect health outcomes.

As for it being completely representative, we strive to match what we know about the demographics of a disease at the stage we are studying it, but it is a complex issue and the research question depends on what

we know about the disease and what we are studying. Therefore, we strive for that. As we move forward and report the results for those groups so we can learn from the outcomes—we know that factors such as sex, gender, race, ethnicity, age, social determinants of health and others all affect health outcomes—we will be able to understand it and next time build a more representative sample, if that information is included and reported transparently.

Q70 **Graham Stringer:** Would it be fair to summarise that answer by saying that to get it theoretically perfect is difficult but that we could get better at it? If so, how do we get better at it?

Dr Clayton: That would be fair, and we would get better at it by being clear about our plans for the target population we are going to recruit for a particular study. We ask for that in our applications; we ask applicants to tell us about the proportion of women and under-representative groups. Now, we even ask for the proportions across the life course: the oldest and old and the youngest and young. Who are they planning to recruit, and why? What is the scientific justification for the target population? We track over time whether they have met their target. Are they close to meeting their target?

We can now track that much more closely since we have examined our inclusion process. We call this inclusion at NIH, and now it looks at sex, gender, race, ethnicity and age. Our programme officers give that feedback to grantees when they are not meeting their targets. We also pause funding if they are way off making their targets, because that is a term and condition of the award. They have to provide a plan in their application as to how they will recruit their target population.

To improve it, we need these steps along the way: a recruitment plan; monitoring; and teeth to the monitoring and the ability to stop funding if somebody is not making their target.

In the end, they need to publish the results in a way that provides information about the evidence of how that treatment or intervention worked in those groups for us to close the loop and have that data to inform the next study. If they include everyone and do not report the results in publications, or in their final progress report to NIH, we have not reaped the full benefits of all those individuals being included in the study. In particular, we know there are some issues for women. As I mentioned, fewer than one third of the large phase 3 clinical studies supported by NIH have had any results reported by sex or gender. That analysis by Geller et al was repeated recently and unfortunately there has been no improvement in that.

Q71 **Graham Stringer:** Where should government focus its attention to tackle the challenges posed by the problem of reproducibility?

Dr Oransky: We have heard some good suggestions along the way here. I would argue, as others have, that it is a multifactorial problem.

Obviously, you have to focus efforts in certain places. I would recommend against focusing on a single area, but there are a couple of key targets. Sanctions are important. Dr Clayton just mentioned the ability to pause and withhold funding for a period of time while a remedy is created.

In the previous panel there was a discussion of clinical trials and whether they were published as they were supposed to be, or at least posted on to sites like ClinicalTrials.gov in the US and around the world. For many years, the FDA did not sanction that, even though it had the power to do that. It has started to do that. This has been reported in the past year, but I would have to look that up to be exact. It has started to make sanctions. That may apply to pharmaceutical companies and biotechs as well as academic institutions that run trials. The sanctions aspect is of critical importance.

I also think that improving rigour means improving incentives. Whether it is the REF or funding mechanisms in general, government can focus on things, or at least broaden the focus beyond—maybe even replace—publication. It is not so much publication of data and the fact that things are public, but publication in certain journals as a metric of what is considered good or rigorous. Anything that gets away from that and incentivises the process and, as Dr Clayton mentioned, includes enough detail about the methods, the participants, etc., would be a positive development.

You have seen that happen. I refer back to clinical trials. This has taken time. I could argue that it really started 20 years ago, but there were different steps along the way. Some were by journal *[Inaudible.]* publish a paper if you have not registered your trial on one of the clinical trial sites like ClinicalTrials.gov.

Then the US Congress said that you must do this, not even as a condition of funding; it went beyond that. If you are doing human subject research, or clinical trials, you must as a condition register your trials in advance.

There are corollaries to that, as you heard from Professor Munafò, about pre-registration and registered reports. The details of that need to be worked out, but there are at least parallels.

I think of this as the upstream and downstream. There are many parts of the stream that should also be targets, but those are two I would highlight first and foremost.

Dr Clayton: I would agree with Dr Oransky that this is a complex issue; it is a research ecosystem issue. In order to address it, we need system-wide solutions and to address it from multiple vantage points at the same time in a co-ordinated manner. Training in this area is critical for future generations, current early-career investigators and senior investigators on some of the issues about statistical design and rigorous design that affect reproducibility.

We have recently added the element of training in rigour to our training grants and career development awards. They will be assessed on how well they plan to train investigators and trainees in that way.

We need policies that represent standards. This is a standard you must meet. Sex is a biological variable and the NIH inclusion policy says you must include women and under-represented groups, unless there is a scientific reason not to do so. You must justify any single-sex study scientifically. Cost is not a factor that can be used to justify that.

As Dr Oransky mentioned, we need incentives, and I agree with that. We need incentives to support research on reproducibility. We need programming to support research on sex and gender. Our office does that in collaboration with the 27 institutes and centres at NIH where we bring attention to these issues and provide additional funding for investigators to look carefully at the role that sex and gender may be playing so we can get more unbiased, robust results reported to inform the entire evidence base, because we know that historically there has been over-reliance on male animals and men.

Our business processes need to change and include the evaluation of rigour and transparency as two key factors related to reproducibility. I refer to peer review of applications and manuscripts; programmatic oversight; and bars to funding. We need practice and process changes, training and policy. We like to say that we need carrots and sticks to initiate the behaviour and culture change that we want to see in science, but working together and addressing the problems from multiple vantage points in a co-ordinated manner and then assessing what is working along the way would allow us to do this carefully and learn from each step.

Chair: Dr Clayton and Dr Oransky, thank you very much indeed for your evidence today. This is the first of our hearings. You have given us some very important directions to follow up in future hearings. As I said at the beginning, we are particularly grateful for the fact that you have joined us so early in the morning. We have seen dawn break in New York over the shoulder of Dr Oransky. Thank you very much indeed for getting up early. I wish you well for the rest of the day.