ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

EVALUATION

Work-in-progress report based on RSS event taking place on 12 July 2022

Introduction

This report documents the discussion at the fourth of the RSS's Covid-19 evidence sessions.[1] Before the session, we identified five questions designed to discover what lessons around evaluation can be learnt – both so that we can learn from what went well as well as reflecting on where there are areas for improvement. We sought to bring in a wide range of views during the discussion at the event, and these are reported below. However, even over a two hour meeting, only so many people were able to speak, so this is intended as a reflection of views expressed during the meeting and should not be read as representing the views of the RSS.

The five questions we explored were:

1. How are risks to successful evaluation – low participation-rates, biased comparators, weak performance-monitoring, delayed transparency – best managed?
2. Which Covid-health policies needed evaluation – not all do and so who decides; and are there some criteria we could set out which might us help prioritise those policies which could/should be evaluated?
3. How were prior beliefs formulated about the likely impact of public-health policies during Covid – via modelling, formal elicitation etc?
4. Medicines and vaccines apart, what are the barriers to randomization when evaluating non-pharmaceutical interventions?
5. What were some of the benefits that evaluation brought during the pandemic? What were missed opportunities? How can statistical science help to deliver more robust policy-evaluation in the next decade?

List of speakers

Main speakers

- **Mahesh Parmar** – Director of the Institute of Clinical Trials and Methodology, University College London (UCL)
- **Theresa Marteau** – Director of the Behaviour and Health Research Unit at the University of Cambridge
- **Tim Peto** – Co-Leader for the Infection Theme of the Oxford Biomedical Research Centre
- **Isabel Oliver** – Interim Chief Scientific Officer, UK Health Security Agency (UKHSA)
- **Chris Robertson** – Professor of Public Health Epidemiology, University of Strathclyde

Contributing speakers

- **Tom Whipple** – Science Editor, the Times
- **Ed Humpherson** – Director General for Regulation, Office for Statistics Regulation (OSR)
- **Julian Peto** – Professor of Epidemiology, LSHTM
- **Arun Chind** – Occupational Health Physician, Consultant Economist & Statistician at Proshen Health & Risk Consulting
- **Nigel Jacklin** – Founder and Managing Director, Think Media Consultancy
- **Tracey Brown** – Director, Sense About Science

[1] This document is a work in progress – a final version will be published in 2023. If you notice errors or omissions please email policy@rss.org.uk.

- **Susan Hopkins** – Chief Medical Advisor, UKHSA
- **Ian Russell** – Emeritus Professor of Clinical Trials, Swansea University
- **Iain Buchan** – Chair in Public Health and Clinical Informatics, and Associate Pro Vice Chancellor for Innovation, University of Liverpool
- **Camila Caiado** – Professor of Statistics, Durham University and Fellow of the Wolfson Research Institute for Health and Wellbeing

1. How are the risks to successful evaluation best managed?

Theresa Marteau suggested two ways in which we could mitigate some of the risk to successful evaluation. First, is by ensuring that the teams who are conducting evaluations have evaluation expertise and experience – these teams should include statisticians as well as design and content experts. Too often during the pandemic this did not happen with policy evaluation. Second, it would be useful to have a set of principles. Below are four principles that were included in a SAGE paper setting out a science framework for opening up group events.

- **Design**: study designs should optimise causal inference, ranging from randomisation for large scale events, to meta-analyses of case control and cohort studies
- **Measures**: there should be a core set of measures across studies – biological, environmental and behavioural
- **Ethical**: studies should generate high quality evidence, transparently, treating everyone with equal moral value
- **Open Science**: the practices of open science should be followed, including pre-registration of protocols

Isabel Oliver set out some of the challenges to successful evaluation. In health protection, we often need rapid evaluations and this can result in pressure that leads to suboptimal evaluations. What is even worse is that, in the past, we have suffered from an approach which focussed on implementing interventions without necessarily considering evaluation. There are examples where antibiotics for treating people with certain bacterial infections have become standard practice without an evaluation being conducted, when one could have been done when the threat was first identified.

One key problem, **Oliver** suggested, is that there are different views on what a successful evaluation is. A medical epidemiologist looking for a robust and reliable evaluation may have different criteria from policy colleagues, who may not be as focused on robustness.

Oliver also noted that for rapid evaluation, we often rely on routinely collected data with all of its limitations – missing data, misclassification, confounding, etc. When conducting trials, participation rates can be low. One of the good things during the pandemic was an increase in participation and engagement with research, eg in randomised trials. **Oliver** also suggested that the reliability of model-based evaluation of routine data is difficult to assess and requires a lot of thought about how to make sure that there is robust information about parameters. One thing that she has sought to do in the past is strengthen the reporting and publication of outbreak investigations in a way that allows the extraction of parameters that can inform transmission models, because often that's where the evidence is coming from.

Some interventions are easier to evaluate than others. There are also some aspects of evaluation where it is difficult to get engagement and interest from others, whereas in other areas engagement is more straightforward. But, to a degree, the same infrastructure and expertise for randomisation used in diagnostic or therapeutic evaluations, can also be used in policy interventions.

Oliver argued that there's been a perception that the need for a rapid evaluation is in conflict with research transparency, openness and reproducibility. During the pandemic, she suggests that UKHSA has shown that that's not the case – but it is important to receive challenge and pressure from academic and other colleagues on this to ensure that it keeps improving. The same applies to patient and public engagement, which can be neglected when there is a need for rapid action.

There are also barriers around data access, though **Oliver** suggested this situation improved over the pandemic.

Camila Caiado looked at preparedness for the next crisis – especially related to how the RSS and wider statistical community in the country should respond. One aspect that was quite problematic from the beginning is that the relationship between statisticians and local authorities' public health teams was close to non-existent. This was especially true for university-based statisticians in England. In Scotland there were better teams and partnerships. Statisticians at Durham University had sought to build a partnership over the past five years. Being embedded as part of the council was incredibly useful. But other local authorities in the region did not have the same sort of support. The RSS was driving a national response while potentially neglecting the local response, argued **Caiado**.

As part of the Durham University service evaluation on the deployment of lateral flow tests, there was a very tight community and good participation rates – It was the build-up on how to engage the community that was incredibly important. Because there was already engagement with the local public health team, the service evaluation could be deployed relatively easy.

2. Which Covid health policies needed evaluation and how should they be prioritised?

In answering this question, **Mahesh Parmar** stressed that we first need to appreciate that many interventions – both health and policy – don't work. When it comes to policy interventions there is a question as to whether there is the political will to evaluate a policy when there's a reasonable chance that any evaluation will suggest that the intervention did not work as hoped. The realpolitik, is that the appetite for this might be lacking.

There are some instances where evaluation makes sense – for example, in the clinical trial space, where there is a well-established way of getting an answer. And there are some areas where you clearly can't do an evaluation anyway. But there is also a middle space where you might be able to do effective evaluation. If we're going to do more evaluations in this category, **Parmar** says, the primary outcome measures need to be short-term. Long-term outcome measures for these policy decisions won't be able either to inform the policy decision or to change a policy which has already been partially rolled out. It is also vital that evaluations are launched quickly, so that they can inform changes to policy in a timely way. So, when we have a policy intervention that is big and uncertain enough to warrant evaluation, we need to ask whether it can be evaluated in a sensible timeframe.

For example, we may think that lockdowns are very difficult, if not impossible, to evaluate in a sensible way in a sensible time frame. Is there a relatively short-term primary outcome measure that will help you make a decision? Is the result of any evaluation likely to be available in the time scale to influence or change policy? If it isn't, then there is little point in evaluating it, aside from getting a retrospective view on whether it was the right thing to do. A retrospective view, whatever the result of the evaluation, is unlikely to be a vote winner.

Parmar highlighted two other important things to consider. First, is the 'signal' likely to be large enough to detect with the design proposed? For example is the RCT required to detect the signal with reliability, or is the signal so large that an observational study (despite the potential confounders) will be sufficient to identify it reliably and robustly? Second, we ought to consider other types of RCT – such as a cluster RCT – which may be suitable for a number of policy decisions.

With those questions set out, **Parmar** suggested that there ought to be a systematic review of the major policy decisions during the pandemic that were evaluated – to understand the extent to which evaluation informed and improved policy. There should also be, he suggested, a review of policies that were not evaluated to consider whether evaluation was possible or desirable. If so, why was the evaluation not carried out? Was it just realpolitik, or were there other issues – for example, around timeframes, resources, expertise etc. If some evaluations that weren't carried out were possible, how might these studies have been designed and run? This exercise would be helpful to learn about how we might design studies during future pandemics.

Parmar stressed that evaluation needs to move at incredible speed and with a great deal of efficiency in this arena. So, we should consider whether it is possible to set up protocols so that for the next pandemic we have some

"evaluation ready" outline protocols. We should also explore whether we can develop platforms in the same way that we have done in clinical trials to evaluate policy decisions, eg on the back of observational studies.

Theresa Marteau suggested some criteria for assessing which Covid health policies need evaluation. First, we should look at spend. Where there's a large budget – for example, for NHS Test and Trace, which had an indicative budget of £37 billion – a robust evaluation plan should be part of that programme. Test and Trace was set up without an evaluation plan (this has also been noted by the Public Accounts Committee) – lots of performance data were published, but it was still not possible to infer an estimate of effectiveness.

Marteau's second proposed criterion is to look at the significance of any policy in achieving its core aims. During the pandemic there were at least two core aims of any policy: reducing transmission, and reducing health inequalities (for example, financial and other support for self-isolation, and tailoring vaccination rollout to optimise uptake in all social groups). In the case of both these examples, there was no robust evaluation plan.

Isabel Oliver suggested that from a public health perspective, most Covid policies needed evaluation – so prioritisation is key. There is a lack of evidence on non-pharmaceutical interventions, but it was also important to evaluate Test and Trace, where enormous amounts of money were spent and where we are none the wiser as to what the best approach for large scale contact tracing is. UKHSA are doing some evaluations on this retrospectively using routine data, but it was challenging to do that in the midst of the pandemic, on a highly sensitive issue, in a politically charged environment.

The pandemic was a very busy period where policies were proposed and developed on a day-to-day basis, which made it challenging in terms of prioritising areas for evaluation. **Oliver** gave the example of an evaluation that did inform policy on time. This was challenging because of how quickly the situation was evolving. Early in the pandemic there was a proposal on the potential for screening parts of the population, particularly healthcare workers, with rapid antibody tests. UKHSA's evaluation highlighted that the positive predictive value in the context of the prevalence that could be expected at the time, would be of concern and lead to significant errors in the reporting of cases. This informed the policy and prevented government spending lots of money on rapid antibody tests.

UKHSA also evaluated the impact of local contact tracing partnerships. They worked with statistical colleagues, using time series to help understand the best approaches for contact tracing. They also evaluated daily testing of contacts of confirmed Covid cases as an alternative to self-isolation. Unfortunately, by the time the results were ready, the policy around testing and isolation had changed.

In terms of how the statistical community can help, **Oliver** highlighted four key areas: working with UKHSA (in peace time as well as times of crisis), helping UKHSA develop methods for rapid evaluation, continuing to provide advice and support in emergencies, and collaborating to develop pre-ready protocols.

The speakers discussed the question of whether the decision to give the vaccine with an interval of 12 weeks between the first and second dose (rather than three weeks as originally licensed) should have been subject to some form of randomised evaluation. **Chris Robertson** suggested that while we would ideally have tried to randomise, in this case it would have been phenomenally difficult to convince everyone the process was fair due to the tight timescales. This may be a case where a retrospective evaluation would help. There's a similar question as to whether there was a difference between the various vaccines. In Scotland different groups of people received different vaccines – eg, if you were over 80 and lived in a care home you got the Pfizer vaccine, whereas if you lived in your own home you received AstraZeneca – and, while this isn't truly randomised, it would allow some retrospective analysis. **Parmar** suspected that there may not have been the political will to answer the question about whether the shorter or longer interval was better – to be ready for a rollout and discover that the other method was preferable would have been a real challenge. **Tim Peto** suggested that rolling out the vaccine by age was very sensible and seemed to work well for uptake – though there was perhaps an opportunity to roll it out differently in some regions to compare the relative merits of the two approaches. There is also a particular challenge around getting ethics approval on these sorts of timescales, and that is worth looking at if we want to be able to move faster. **Peto** also noted that there may have been concerns about whether Pfizer's claim that the

doses should be given three weeks apart may have been driven to some extent by commercial considerations. **Marteau** wondered what the scientific uncertainty was and what the priors were for going for three versus twelve weeks – if one saw the uncertainty and the scientists were pushing for it, maybe something would have been possible.

Tom Whipple examined the public debate around masks. As a journalist, he has keenly observed what happens when there was public debate with a lack of evidence to inform it. He was invited onto the radio to discuss masks. The presenter asked him to outline the evidence for masks, expecting him to slap down anti-vaxxers. But he had to pause and say the evidence wasn't great – which was a difficult thing to say, given the debate had become political. Part of the reason for this is what has happened with the evidence – it has become politicised. People can point to portfolios of evidence on masks, but when areas aren't politicised one study is enough). The problem was that while masks probably do provide some benefit, the evidence base wasn't provided until too late. **Whipple** pointed to two decent trials. One was the Danmask study, which looked at the effect on the wearer but couldn't find effect beyond about 50%, which would have been a huge finding. There's also a Bangladeshi cluster RCT, where masks were encouraged in some villages and not in others. In that trial they found that surgical masks work pretty well and cloth masks work less well. However, this is the only trial **Whipple** is aware of, and it was conducted in a very demographically and socially different place, with a completely different climate. **Whipple** thinks that lives might have been lost because by the stage study results came out, the issue was so political, the results didn't matter. Given the amount spent on other areas, investment in providing better evidence earlier on here could have had a huge positive benefit.

Julian Peto discussed evaluation of Test and Trace. He argued that if you want to test and trace it has to be "an assiduous test" and you need to test the whole population regularly to have any chance of picking up a substantial proportion of cases that you wouldn't pick up anyway symptomatically. This would also mean being in touch with people so that you can give them advice. To test the whole population once a week (which is the sort of frequency you'd need to be effective) would require ten million tests a day. NHS Test and Trace, with £37 billion and vast numbers of Deloitte consultants, got up to 800,000 PCR tests at its peak, which is less than a tenth of what was required to have a serious impact.

One thing that was promised from the very beginning of the epidemic, and publicly announced in April 2020, was that the entire population of Southampton was going to be tested once a week. That trial was never done and that's the one evaluation which **Peto** thinks would have been especially important. The point about testing the whole population is, if it is done by an NHS lab, it goes straight onto your NHS record so you would have real time information about who's getting infected and how infection spreads in locked-down households. What was done eventually was spraying the population with lateral flow sets which were self-administered, very often not reported and certainly less sensitive than better methods. It would help with future pandemics to have facilities in place that would allow the immediate testing of the whole population once a week. This could have controlled the pandemic without the necessity of a lockdown.

Tracey Brown discussed the transparency of evidence framework that has been developed by Sense About Science and now features in the Treasury Green Book. Confronted with the first instance of mass engagement with the policy interface, Sense About Science has just completed an inquiry, called "*What Counts?*". For the first time, they were able to survey the population as to their experience of trying to get hold of policy evidence. They ran a NatCen survey last year, interviewing everybody from rural bus drivers to fostering agencies about their experience of trying to implement policy and get evidence. The key point is that while it's all very well to dig down into trials, unless we have honesty about what the policies goals are, we can't really ask good questions about what we're seeking to optimise. That's been one of the biggest problems, all the way through: modelers were not asked where's the sweet spot between closing schools and preventing transmission. If they had been asked that question, they would have probably come up with a fairly good answer, especially since Warwick University were involved in the modelling, and they actually specialise in these kind of questions. There was a great dishonesty with the public and a real sort of squeamishness about talking about the fact that there were trade-offs to be made and what we were looking for was how to be the most effective at the least cost to society.

This, **Brown** suggested, impacts on the question about how we prioritise what is evaluated. There needs to be consideration of the dilemmas that confront different people. It is not experts, media, policymakers and then just "everyone else out there". What is "out there", is thousands upon thousands of people having to make complex risk assessments. For example, in "What Counts?" bus drivers were wondering things like "how do I compare the risk of leaving a kid at a rural bus stop with no pavement, because they haven't got a mask, with the risk of letting them on the bus?" It's two different types of risk. Fostering agencies were wondering things like what to do when a child won't put a swab up their nose because they're being hysterical about it. Do they let the foster placement break down? That's a nonlinear risk. It breaks. It's ended. The child's devastated. How do you weigh up those situations? It's those kind of trade-offs that thousands of people across society were forced to make. And many of them felt ill-equipped. So whenever we look at what we're going to do next, we also have to think about the mechanisms, both political and research-based, that solicit those questions, and draw in those dilemmas very quickly. Especially when you're asking masses of people to overturn the services and facilities that they provide and to provide them differently. To make that kind of trade-off, we need their input back really quickly to decide how to set up the evaluation question in a sensible fashion, and to ensure that it answers the relevant questions.

Susan Hopkins argued that, in preparation for the next pandemic, we need to think through very clearly what we want to evaluate, why we want to evaluate it and how those evaluations are going to be done. It is extremely challenging to conduct studies at rapid pace with rapid results in a way that can inform policy. The only way we can do this effectively is if we have thought through the possibilities before, have things ready and have government engaged with how those studies work and what we could do and should do in an emergency situation. The ideas that came through in many areas could not be evaluated at pace, because of policy decisions that were being made. Some of those were evaluated retrospectively, but those evaluations are never going to be as effective and as robust as prospective, thought-through, open science papers, with open science protocols. One of the learnings from this pandemic is thinking through the possible interventions, the possible evaluations that we can do and having those ready and waiting for the future state.

Ian Russell discussed a symposium he organised on 14 April 2021 for the Royal College of Physicians of Edinburgh, on the public's health after Covid. The symposium included discussion of eight evaluative papers, seven of which were quantitative and one qualitative. **Russell** edited those papers for publication in the supplement to the June 2021 issue of the College Journal. Both the symposium and supplement were mainly statistical. Though the editorial team judged all papers as high in quality, most of the compliments received mentioned the sole qualitative paper. This carefully reported six case studies: four about homeless people receiving special help during Covid, and two about successful service changes in response to Covid. This preference for individual stories reflects the reported views of the media and the public. Triallists face similar dilemmas. Qualitative analysis of subgroups often gets more attention than full statistical analysis of the trial population. Trials units have responded by developing standard operating procedures to enhance trials by adding qualitative flesh to statistical skeletons. Statisticians who accept the public's preference for stories over robust analysis need to work with rigorous qualitative researchers to integrate their evidence with ours, argued **Russell**.

3. How are prior beliefs formulated about the likely impact of public-health policies during Covid?

Theresa Marteau suggested that prior beliefs were formulated using a mix of methods, of which modelling and formal elicitation are two. We could also perhaps add to that, rapid reviews of existing salient evidence – including evidence-synthesis where possible. During Covid, there was a tidal wave of evidence and it is important to build on the methods that have been developed during this period. This means using automation, crowdsourcing and single databases (eg, OpenAlex) to inform policy in and outside of emergencies.

Sheila Bird discussed the elicitation of priors, giving the example of Operation Moonshot, where the policy-makers seemed to think that three quarters of the public would turn up to be screened weekly using lateral flow tests – when, in fact it was only a quarter. No experienced biostatistician would have imagined that they would get near 75% -- and so there is a question as to how the elicitation of priors was conducted. **Marteau** suggested that there are priors that might exist in policy circles, but SAGE papers are full of priors which are based on estimates of

people's behaviour in similar contexts. There are different methods, and it is possible that there is political motivation for some methods. However, the papers generated in SAGE were using a proper mix of methods and estimating a range.

Nigel Jacklin discussed lockdowns and low threshold case definitions. He is not convinced by the official narrative that Covid was a very deadly disease whose impact was mitigated through exceptional and heroic measures. His alternative hypothesis is that Covid was a fairly bad disease and the measures taken did more harm than good. He questioned to what extent the decisions about aspects such as lockdowns and the case definition were based on political judgment, or and the extent to which decisions were influenced by a narrow field of experts with a predetermined view.

4. Medicines and vaccines apart, what are the barriers to randomisation when evaluating non-pharmaceutical interventions?

Theresa Marteau highlighted two connected barriers to randomisation. The first was that evaluations, certainly policy evaluations, have not been conducted by teams with evaluation expertise – some teams didn't have statisticians or design or content experts. That reflects a weak and failing system of policy evaluation across government, which predates the pandemic. There is good government guidance in the green and magenta books and those are regularly updated. But that guidance most often is not followed. The NAO in December 2021 estimated that only 8% of major spending projects had a robust evaluation plan built in. The NAO has previously highlighted a lack of political engagement in evaluation as well as a low capacity for delivering it.

Tim Peto discussed the background and challenges associated with an important cluster randomised trial he undertook, which looked at daily contact testing versus self-isolation for school-based Covid contacts. In January 2021 pupils were being screened and the contacts of positive individuals (in pre-determined bubbles) were quarantined for ten days – disrupting the functioning of schools. The Department for Health and Social Care (DHSC) and Department for Education (DfE) approached **Peto** to ask him to undertake a trial to determine whether a policy of daily contact testing with a lateral flow device with the release of negatives was safe and effective, with a view to report the results by end of June 2021.

There were 201 schools involved in the trial – randomised to 99 control schools and 102 intervention schools. There were two primary questions that this study was looking to answer. First, is it safe - what is the rate of symptomatic infection in control and intervention schools? Second, does it improve attendance – what is the rate of Covid-related absences in the two groups? As a secondary outcome, the team also looked at how often contacts actually got the disease.

Initially they had sought to get data from schools – but schools were not used to providing data, and it proved to be unreliable and incomplete. The approach had to change halfway through, and data was collected directly from DfE instead. The team were also able to get some idea of the immediate contacts of index cases by using PCR tests for research purposes. Pupils were asked to do two PCR tests – one at the start and one either on testing positive or, if their lateral flow tests were always negative, on the last day and they did not get the results until the end of the outbreak. This allowed the researchers to assess what happened without intervention. This research just needed individual consent.

The results of the study showed that in secondary school, college students and staff, infection following contact with a Covid-19 case at school occurred in less than 2% of cases. This meant that there was no evidence that switching from isolation at home to daily contact testing increased rates of symptomatic Covid in students and staff. The study suggests that daily contact testing is a safe alternative.

There were a number of challenges in conducting this trial. It was conducted in 'pandemic time' meaning that it could not be peer reviewed in the normal way. There wasn't time if it was going to have impact – the departments did not even consider peer review. As it happened, the DfE wrote the first draft of a report, which **Peto** himself

reviewed. DfE generally felt a strong sense of ownership and found it difficult to let go of control of any aspect of it. Another big problem was that no clinical trial centre could provide facilities at short notice.

Peto also highlighted that neither DHSC nor DfE had much experience of clinical trials, which also introduced challenges. In particular, they did not seem to understand the need for a control group – the analysis team in the DHSC had never considered how to present results for both control and intervention groups. They also didn't appreciate the "intent to treat" principles, so the people who got consent from schools didn't inform them that they'll get randomised and didn't explain how dropping out would impact the results. Around a quarter dropped out in the end. It was a commercial company that went to schools to secure their participation and **Peto** didn't appreciate the related difficulties until the end of the project.

There was also, **Peto** argued, an issue with schools and data. He had thought that schools would have the same knowledge of data handling as hospitals but there was certainly no equivalent of a local trial physician who understood trials and the importance of obtaining data.

Finally, there were particular issues with this trial due to its political salience. Politicians were highly interested in the results and they were asked every week to give even just a hint of the results. There was antagonism from parents, staff and public health doctors against direct contact testing (DCT) but, by contrast, some schools withdrew because they were not randomised to the DCT arm of the trial. There was also a lack of confidence in the PHE Ethics Committee, and concerns around understanding the ethics of a cluster randomised study where parents weren't being asked for explicit consent to expose their children to perceived increased risk.

To address some of these challenges, the team created an Independent Trial Steering Committee which met weekly. Sarah Walker was critical behind the scenes. The appointment of a trial statistician with good domain knowledge who worked with DHSC data analysts was also crucial. **Peto** also argued that it was important to keep government departments at arms' length to protect the trial from political interference.

5. What were some of the benefits that evaluation brought during the pandemic? What were missed opportunities?

Theresa Marteau highlighted some of the benefits of the events research programme, which was the largest non-pharmaceutical policy evaluation during the pandemic. This research estimated the risk of transmission at large scale events ranging from football to opera. The chief science advisors at the Department for Culture, Media and Sport (DCMS) and the Department for Business, Energy and Industrial Strategy (BEIS) played a vital role in setting up this evaluation.

For this research, carbon dioxide was recorded in multiple places at live events as a proxy for ventilation. This allowed a quantification of the variation in ventilation within venues. The maximum and average level of carbon dioxide measured at various venues is shown below:

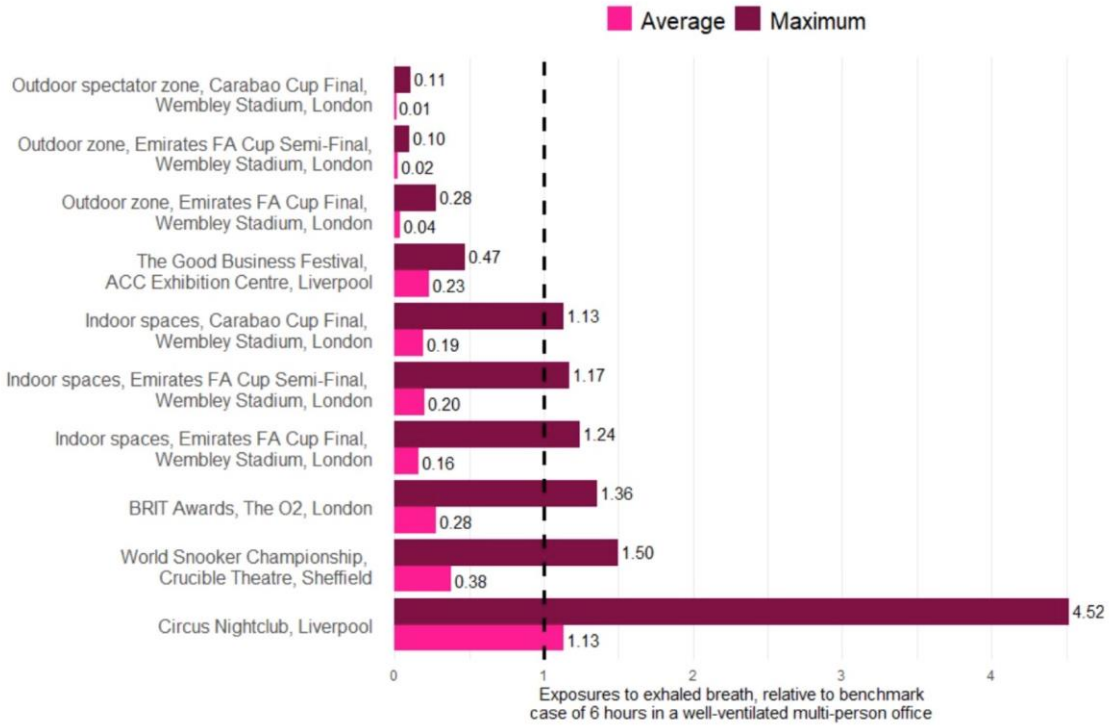

Image from DCMS's Events Research Programme: Phase I findings

There is a difference between indoor and outdoor venues – the average and maximum levels measured at Wembley Stadium are relatively close. For indoor areas there is a difference between the average and the maximum, which tells us that levels change substantially at different stages of events in those venues.

Marteau also identified some missed opportunities. First, there was the opportunity to evaluate the impact of financial support for self-isolation in the lowest income households. SAGE advised, in the middle of September 2020, that there should be urgent provision and evaluation of financial support to avoid loss of income in the poorest households. This would help to realise the considerable investment in NHS Test and Trace and help to prevent disease, as well as contribute to economic recovery. However, the government introduced a different intervention – giving a fixed amount of £500 to those eligible for benefits and fines of up to £10,000 for those not self-isolating. The impact of this wasn't evaluated – but it is reasonable to think that, if people were worse off because they were having to self-isolate, these measures may not have been effective in encouraging self-isolation.

The other big opportunity that **Marteau** highlighted was around the vaccination programme. Based on what we know from other vaccination programmes, SAGE (with high confidence) predicted lower uptake in some social groups. The government rolled out the vaccination programme without an evaluation plan to, eg, compare different approaches to rollout to mitigate the predicted lower uptake for some groups. As of April 2022, ONS figures suggested that the mean vaccination rate (for three doses) among the population as a whole was 74%. But this ranged from 59% in the most deprived quintile to 89% in the least deprived quintile. Had different interventions been evaluated, it's possible that the gap could have been smaller.

Marteau suggested that to tackle this, statistical science should work with GO-Science and the policy profession to strengthen the weak existing system of policy evaluation for all three stages of policy-making: policy design; monitoring during implementation and adjustment when off-course; and outcome evaluation. This should happen both inside and outside of emergencies. To effectively strengthen this process, there need to be situations in which evaluation is *required* – ie, evaluation cannot just be recommended as part of guidelines. It is also important that any evaluation is published transparently so that it is open to public scrutiny, which will lead to more effective policies.

Chris Robertson detailed a collaboration that he was involved with in Scotland – the Early Vaccine Effectiveness study or EAVE II. There was a large team across academia, operations, public patient individuals, project managers, data analysts, and many people who worked within Public Health Scotland (PHS). All of the analysis with done within PHS, using honorary contracts or by directly employed analysts.

This collaborative work started about 12 years ago when the collaboration was looking at early assessment of vaccine effectiveness during the H1N1 influenza pandemic. The team recruited individuals from 40 practices and looked at the data within PHS. They carried this work on to look at seasonal vaccine effectiveness. Their first study took two years to complete, and was published in 2012 – a couple of years after the H1N1 pandemic. A separate study on seasonal flu upped the sample size to 230 GP practices (about a quarter of GP practices in Scotland) and carried out the work in the NHS Scotland EDRIS Safe Haven. The Cabinet Secretary for Health and Sport (Jeane Freeman) asked the team to scale up to the whole Scottish population, which was possible due to a partnership with Albasoft who were able to extract real time data from GP databases and feed it into PHS.

There were three aims for EAVE II:

1. To evaluate the effectiveness and safety of vaccines.
2. To understand the epidemiology of Covid-19.
3. To explore different patterns of healthcare utilisation and outcomes.

Robertson emphasised that the data structure was important for the success of the project. Because they had ethical and privacy permissions to access data from everyone, this gave them a population spine of everybody who was living in Scotland at the beginning of the pandemic. Through accessing all of these routinely collected healthcare datasets, they could track people through the pandemic. There was data linkage across hospital admission data, primary care consultations, prescribing data, mortality data, lab data, vaccine treatment data, self-reported data (from the Census and Test and Protect Surveys), tele-consultations and birth and pregnancy data. These were all linked together within PHS at various times for this analysis.

EAVE II has had quite a lot of major outputs, of which **Robertson** highlighted two. First, they assessed real world vaccine effectiveness of one dose of the vaccine in Scotland. The results started coming in around February 2021 showing the effectiveness of the vaccines. The key thing about this piece of work was that the same analysis was conducted in multiple ways within the same data set. This was also the case when the team looked at the safety studies on thromboembolic events in Scotland. They did not just use their self-controlled case series, but were also able to do case control studies using this mechanism within their cohort. This was followed up with colleagues in Brazil looking at the same analysis with the same vaccines.

Robertson suggested that the success of EAVE II was based on early consultation with general practice representatives from the two main bodies in Scotland. The GPs were happy for their data to be used in this instance, provided they were used only for coronavirus pandemic work. Their main concerns were about patient privacy, which meant that EAVE II couldn't get the full record for a patient, instead accessing only groups of codes together. For example, you might know that a patient had had asthma, but not the specific type of asthma they could have had. The team also had to use an anonymous identifier – different to the identifier used for all of the linkage work within PHS (the Community Health Index). This meant an extra layer of privacy so that it was not possible to go backwards and identify patients.

The second key contributor to success, **Robertson** argued, was the speed allowed by doing the whole project within PHS's Safe Haven rather than within the EDRIS Safe Haven. That sped up the whole analysis greatly. But it meant that there were some difficulties for PHS, who had to set up two confidential areas: one that only the named PHS analysts could access for data linkage, and a separate area that only EAVE analysts could access, but where all PHS data and GP data could be linked. Because they were working within PHS, that meant the team had real time data on a daily basis during the pandemic and it gave them speedy approvals through the Privacy and Public Benefit Committee. It also helped that the projects' principal investigators were all involved in the national committees (eg, SPI-M, JCVI etc) and this meant that when questions appeared they could be looked at very quickly and results could be fed in with enough time to inform strategy.

Having academic researchers working within PHS also meant that they were effectively training the analysts in more up-to-date statistical techniques. But at the same time, the analysts in the universities began to understand the data structures much better – which helped statisticians to be more effective.

Ed Humpherson discussed how effectively official statistics served the public good, in the context of evaluation. A key aspect of statistics serving the public good is that they enable evaluation of policy. There are two senses of evaluation. The first is a broad, almost 'popular' sense of evaluation, and the second is a tight, rigorous, closely defined version. The broad version is evaluation or statistics that enable all kinds of users to form judgments to get a sense of what's going on, to form a picture. And the tight, rigorous sense is focused on quantified causal conclusions based on very rigorously defined methods.

Humpherson argued that in the broad sense statistics performed pretty well. There was lots of data. Those data were often available in granular ways and we know they were used heavily. However, in the more rigorous sense, the results were mixed at best. There are, of course, some good examples of RCT designs and so on, but there was the general problem that you see repeatedly of government of being too programme-focused – ie, focused on programmes as mechanisms for delivering government ministerial priorities and gathering as much data as possible. This, he argued, is true of Test and Trace. It's true of lateral flow tests. It was also true with some of the vaccine roll out. He suggested that the message here for the RSS is that there is a downside for a heavily datafied approach to policy. An obsession with data in policy-making can lead to just gathering numbers for their own sake. **Humpherson** recommends celebrating the ability of statistics to enable lots of members of the public to engage with aspects of the pandemic, and challenging the government to build in evaluation at the outset of programmes. That is certainly what the Office for Statistics Regulation will continue to do, he says.

Arun Chind discussed how behavioural economics could have informed policy. His particular area of concern was that the emphasis was more on passive measures that people were told to do rather than on how they could themselves participate. What has not been emphasised enough is that people should have been given charge of their own health and of their own ultimate destinies. We know that obesity and smoking, for example, have huge implications for all respiratory conditions and when news of the pandemic broke we knew that getting people to slim down and to give up smoking would help them. It would have been an ideal time to drive that message home. This is the nudge factor – and not acting on it means we have lost an opportunity to achieve long-term benefits associated with improving public health.

Iain Buchan discussed the evaluation of the community testing pilot in Liverpool, which he led. The perspective he gave is one of wearing two hats for measuring uncertainty. His first responsibility was as a public health physician, reporting back to his local population for situational awareness and tactical intelligence as the pilot was implemented. The other responsibility was in developing policy evidence that was externally relevant. None of the public health community expected for half of the population to come forward for testing every week. But we thought that this would be a complex intervention surrounding communication and that there would not be a one-to-one relationship between test and transmission chain. 57% of the population from November 2020 to April 2021 came forward for testing. In the initial push with military assistance in the first five weeks there was an overall effect of around a 43% reduction in hospitalisation. Case detection rose by around a fifth and cases decreased by around a fifth. There was then a systematic rollout to the surrounding population, where synthetic control analysis, **Buchan** estimates, led to around a one third reduction in hospitalisation. That analysis was only achievable with all-England data available nine months later.

For the work that the pilot did, which provided tactical intelligence on the ground, the team were able to perform rapid analyses and inform tactical decisions quickly. For example, analysing small areas showed that internet user classification was a more potent predictor of low uptake than material deprivation. This was important because it highlighted how critical digital access was to testing, and affected the rollout of the testing in Liverpool. The national results that the team were able to produce the following summer required access to all-England data, with hospital episode statistics linked to test results. An important lesson from this evaluation, which did have powerful results from synthetic control analysis, was that those policy-relevant pieces of evidence could have been produced very early in 2021, if the all-England data were available at the same granularity and timing as the combined intelligence system that was available in Liverpool.