Written Evidence Submitted by Simon Wood, Professor of Statistical Computing, University of Edinburgh
(C190125)

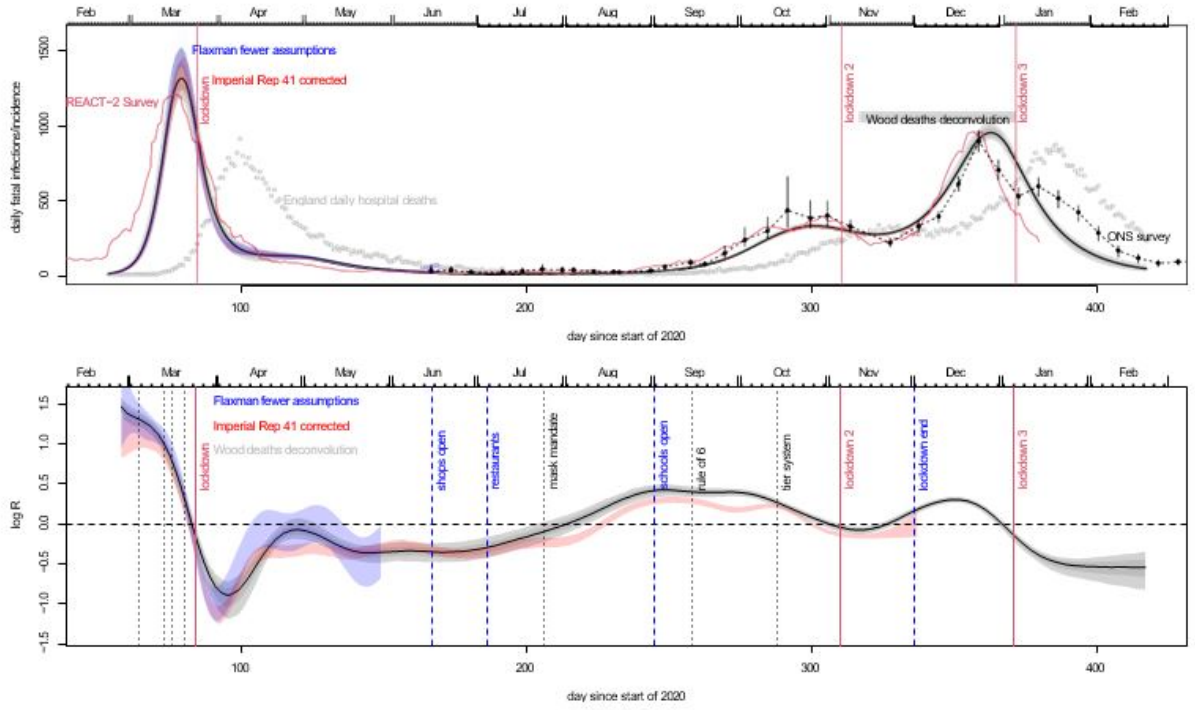
I am writing to bring the committee's attention to some issues related to your 2nd March hearings on Covid statistics and modelling. I write from the unusual perspective of someone whose academic research career started with the sort of biological modelling undertaken by SPI-M, but now working in statistics[1].

The main points I would like to make are as follows.

1. There was evidence from early May 2020 that something other than the 24th March lockdown mea- sures had turned around the first wave of infection, with behavioural changes preceding lockdown being an obvious explanation. This information was available to at least some members of SAGE at the time, but appears not to have been acted upon, and was not well received by the epidemi- ological modelling community. The evidence was sufficient in May 2020 to suggest that something was probably wrong with the modelling based narrative that lockdown was responsible for halting the first wave. By early April 2021 further evidence had accumulated that appears to bring that conclusion beyond reasonable doubt, while showing that a similar story unfolded for the subse- quent two lockdowns. I think that the inconsistency between this evidence and the assumptions underpinning SPI-M modelling should have raised alarms much earlier.
2. There is an obvious logical inconsistency in describing model output as providing 'scenarios' not 'predictions' but then treating those 'scenarios' as 'evidence' to use in decision making. How can a policy maker base a decision on a scenario unless they treat it as a prediction about reality? SPI-M does not in fact present all scenarios as equally plausible, as perusal of their December advice on Omicron makes clear. But it is true that they do not associate probabilities with scenarios. What was not clear from your hearings was that not assigning probabilities is less about choice, than about the absence of any reasonable basis for doing so. Most of the models used by SPI-M have not been validated for prediction. But without such validation there is little reasonable basis for attaching probabilities to the outcomes predicted (if scenarios are not predictions it is not even clear what their 'probability' would mean). Before the pandemic I do not believe that these models would have viewed as adequate for the sort of quantitative prediction for which they have been used (I am not alone in this view, although there is unwillingness to speak out publicly). Treating the models as better reflections of reality than was justified by any data on their predictive performance has led to model results being given unreasonable weight relative to data, in ways that have been unfortunate, particularly in judging the proportionality of lockdown measures.
3. Something has gone awry with the operation of the scientific process over the last two years, with the acceptability of Covid science sometimes appearing to have more to do with its coherence with the accepted narrative than with the objective quality of the work. This appears to me to have led to some weak studies supportive of the accepted narrative being published rapidly in good journals, whereas decent studies questioning that narrative would be repeatedly rejected without review, subject to excessive delay and held up for reasons having little or nothing to do with the substance of the study. In my view this 'groupthink' led to an imbalance in which the benefits of interventions have been overstated, overconfidently, while associated harms have been downplayed unreasonably. This has not provided a good basis for policy.

[1] e.g. the external examiner of my PhD thesis wrote the seminal paper providing the general mathematical definition of the *R* number – I am currently winding up a 4 year stint as joint editor of what is widely seen as the top statistics journal.

4. The sentiments expressed in the hearings, that code should be available for replication and data publicly available, are of course excellent. However in practice this openness has been lacking in key areas. For example, Swiss based colleague Ernst Wit and I were able, with considerable effort, to replicate the Imperial College Covid Response Team's report 41 on lockdown efficacy, based on Imperial's documentation and code. But the same was not possible for either the Warwick or LSHTM modelling that formed the basis for the advice on Omicron. Similarly while the ONS and NHS have made a great deal of data publicly and readily available, key data required to scrutinize SPI-M/SAGE modelling results is restricted to particular modelling groups. One example is the CHESS based information on the distribution of times from first Covid symptoms to death.
5. As your witnesses emphasised, good statistical practices are key to reliably extracting information from data, but it is not clear that SPI-M's expertise in this area matches their level of expertise in constructing epidemiological models. The contrast between Prof Medley's comments on cases data and the ONS data, versus those of your statistical experts, perhaps made this clear, but there are other examples, where inferences are attempted that the data can simply not support. Particularly concerning are attempts to estimate the R number from cases data, an exercise that can not disentangle changes in testing behaviour, capacity and uptake from the changes in disease transmission rates.
6. I was surprised that your witnesses asserted that excess deaths provide a truer picture of UK deaths from Covid than the ONS or government figures. As the largest study of excess mortality over the pandemic[2] is at pains to point out, it is not possible to separate deaths due directly to Covid from deaths brought on by health service and other societal disruptions in response to the pandemic. For example, in the UK context the 13000 non-covid excess deaths recorded in the first 6 weeks of the first lockdown, or the excess deaths at home figures, both suggest that directly linking excess deaths to Covid deaths is difficult.



[2]Wang et al. 2022 The Lancet https://doi.org/10.1016/S0140-6736(21)02796-3

Point 1 relates to apparent discrepancies between model based conclusions and reality, that occurred well before the predictions made in 2021 in July, October and for Omicron. The above figure plots information from five reconstructions of incidence (new infections per day), based on the most reliable data publicly available, but differing in the extent to which epidemiological models are used to extract information from the data, and in the complexity of the models used. The top plot shows the time course of incidence and the bottom the corresponding R number, for the studies for which this is available.

The grey credible bands are from a reconstruction of (fatal) incidence and R from NHS daily death data (shown as grey circles) and published information on the distribution of time from infection to death. The red credible bands are from replication[3] of the modelling analysis in Flaxman et al.'s 2020 paper in *Nature*[4], but with the way that R changes inferred from the data, rather than it being assumed that R changes (instantly) only when government policy changes. The blue credible bands are from a replication of Imperial College Covid Response Team report 41[5], which infers R and incidence from daily data on deaths, hospital and ICU occupancy and testing data. Again the replication[6] allowed the data, rather than model assumption, a greater role in determining how R changed over time. Discrepancies were also corrected between Report 41 and the literature cited as the source for key parameters.

These analyses all imply that incidence was in decline and $R < 1$ *before* each lockdown, implying that the lockdowns can not reasonably have been the cause of the reversal. In addition there are two more direct independent sources based on properly representative randomized surveys, both of which confirm the results. The slightly jagged red curve on the upper plot is from the REACT-2 survey. Those among the survey's random sample who had Covid antibodies, were asked when their symptoms started, in order to obtain the time course of newly symptomatic infections each day[7]. Shifting this trajectory back by 5 days, to allow for the literature based 5-6 days from infection to first symptoms, gives a reconstruction of new infections per day, as shown. Separately, the ONS surveillance survey[8] published incidence estimates based directly on its data, and these are shown as black circles with confidence limits. The ONS reconstructions are not available for the first wave. The various incidence reconstructions are on different scales, so for plotting have been scaled to match on the scale of 'daily new infections that eventually proved fatal'. The different sources agree that incidence was declining well before each lockdown, and suggest $R < 1$ before each lockdown. The first version of the analysis based on daily deaths was provided on a pre-print server in early May 2020[9], was sent to a contact on SAGE and received some press coverage. In my view it was a relatively simple low assumption sanity check of the prevailing narrative on the necessity of lockdown, that should have raised questions as to the reliability of modelling based beliefs about the role of lockdown.

The arguments against the conclusions from these relatively high quality data appear to rest on informal reasoning or on convoluted analysis of much lower quality data. The main informal counter argument is that all over the world we saw a pattern of increasing reported cases and deaths followed by lockdown followed by decrease in reported cases and deaths. Intuitively this sequencing seems to provide strong evidence for the central role of full lockdowns in reversing infection waves. But consider how a different sequence could have occurred. For cases and deaths *not* to have been increasing prior to lockdown a government would have had to decide to lockdown when cases and deaths were already decreasing. Since this would not happen, the only way for lockdown not to precede a decline in cases and deaths would be for cases and deaths never to decline, which is clearly impossible. In other words the observed

---

[3]Grey and red analyses are from Wood, 2021, *Biometrics* https://onlinelibrary.wiley.com/doi/10.1111/biom.13462 [4]Flaxman et al. (2020) *Nature* 584, 257-261 https://www.nature.com/articles/s41586-02002405-7 [5]https://www.imperial.ac.uk/mrc-global-infectious-disease-analysis/covid-19/report-41-rtm/

[6]Wood and Wit (2021) *PLOSONE* https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0257455

[7]Ward H et al (2021). REACT-2 Round 5 https://www.medrxiv.org/content/10.1101/2021.02.26.21252512v1

[8]https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/bulletins/coronaviruscovid19infectionsurveypilot/7may2021#number-of-new-covid-19-infections-in-england-wales-northern-ireland- and-scotland

[9]see https://arxiv.org/abs/2005.02090

sequencing is inevitable *whether or not* lockdown caused the eventual decrease. What matters, for judging whether lockdown *might* actually have caused infections to begin their decline, is what the new infection rate was doing. And new infections happen one to several weeks before detection as a case or death. Other approaches that apparently indicate that lockdowns were essential are based on estimating R from reported case data. As your witnesses made very clear, the reported case data do not measure prevalence (or incidence), and are not even related to it in a manner that is consistent through time. To suppose that such data can none the less be used to infer a quantity related to the rate of change of prevalence is statistical wishful thinking. Then trying to correlate changes in these estimates with containment measures, and treating the relationships you find as representing causal reality is not sound science. The remaining case for lockdowns appears to be that the models say that they are effective, which they do, because lockdown efficacy has been built into their assumptions.

On point 2: When formal validation of Covid predictive modelling has been attempted the results are sobering. See Bracher et al. 2021, Nature Communications, for example[10], where predictive performance one to four weeks ahead was disappointing relative to a very basic statistical baseline model.

The performance of models whose predictive performance is asserted on the basis that they attempt to represent the real mechanisms of transmission is unsurprising, when the detailed descriptions of mecha- nisms implemented in the models are examined. For example, many of the models divide the population (nationally or by region) into 5 year age classes, and then attempt to describe how often people in one age class meet others in different age classes, leading to possible transmission. It is common practice to base these mixing rates (albeit with modification) on the POLYMOD[11] survey data, which asked about 1000 UK residents to keep a diary of their contacts for a day in 2005/6. 7 of the participants were over 75 and none were over 80, meaning that there was no direct data for the age groups contributing half the Covid deaths, even if you believe that the data are otherwise adequate to the task for which they are being used. Many such assumptions and simplifications, based only loosely on data, are compounded within the models. For example, both the Warwick[12] and LSHTM[13] models, used to produce Omicron scenarios, employed assumptions about vaccine efficacy against hospitalization with only weak support from data. The wide range of 'severities' considered in the Warwick modelling was supposed in part to allow for the uncertainties in this assumption. But the Nature medicine paper[14] that LSHTM used to obtain their figures for vaccine efficacy against hospitalization suggested uncertainties in this assumption so wide that essentially anything might happen. This uncertainty was not propagated into the modelling.

While the compound effect of numerous weakly-tested or un-tested assumptions within the models ought to suggest extreme caution in treating their scenarios as relating to possible realities, there was also direct evidence that calibration might be poor quite early in the pandemic. As Dr Ali pointed out, if we had locked down last December, infections would have declined much as they did, but this would have been attributed to the lockdown. The data in the above plot suggest that something similar in fact played out when we did lockdown during the first pandemic year, and the apparent success of the model predictions of declines *because* of lockdowns was equally illusory. In addition there is the difficulty of Sweden. Imperial released predictions that Swedish policy would lead to somewhere around 40000 first wave deaths[15], which obviously did not happen. A since withdrawn pre-print using the Imperial code to makes predictions for the Swedish policy suggested a catastrophe that never happened. The Flaxman et al. paper, discussed above, purported to show that lockdowns were essential to turn around waves of infection across Europe. But their model was forced to treat the final Swedish intervention *as if it was lockdown*, to explain the Swedish data in this way.

---

[10] https://www.nature.com/articles/s41467-021-25207-0

[11] https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.0050074

[12] https://www.medrxiv.org/content/10.1101/2021.12.30.21268307v1

[13] https://www.medrxiv.org/content/10.1101/2021.12.15.21267858v1

[14] https://www.nature.com/articles/s41591-021-01377-8

[15] https://www.imperial.ac.uk/mrc-global-infectious-disease-analysis/covid-19/report-12-global-impact-covid-19/

On point 3, let me recount my own experience. The reception of my paper inferring incidence trajectories was somewhat negative. Versions of the paper were rejected by several journal editors without being sent for review. I eventually sent it to *PLOSONE*, where the only criterion for publication is supposed to be whether the paper is technically correct. After a wait of months, it was rejected on the basis of a report that stated, in essence, that although the analysis and results were correct it would be irresponsible to publish it, as a reader might interpret the results as suggesting that lockdowns were not necessary. An appeal 'succeeded' only for the paper then to be held up for months by an editor who objected to it without finding an actual technical fault. I eventually withdrew the paper and sent it to *Biometrics*, where pre-pandemic academic standards still applied and it was refereed and published rapidly (no editor or reviewer ever suggested that the result or method were incorrect). The replication of Imperial report 41 fared somewhat better, but was still subject to months of delay. Referees did not find fault with any of the conclusions or technical content, but did seem to have problems with the suggestion that lockdowns might have negative consequences that should perhaps be considered alongside any benefits, and with the general notion of critically evaluating others' work. As disturbing as the obfuscation and delay in the peer review process, was how the papers appeared online. For much of 2020, the Google Scholar record for the first paper pointed not to the arXiv preprint server where the paper was posted, but to the extreme right wing 'Daily Stormer' website. Google do not normally index this site. fullfact.org referred to the study in an article on lockdown efficacy, giving an entirely misleading account of why its results differed to those of Flaxman et al, while also getting wrong basic information from two papers on the distributions of times from onset to death. They declined to correct these misrepresentations, nor the knock on problems they caused for their own misleading analysis of lockdown efficacy. The second paper was simply not indexed by Google Scholar or Web of Science for 5 months post publication, unlike every other paper in the same or more recent issues of the journal that I checked. Eventually this was corrected. From talking to other people who have tried to publish papers that went somewhat against received opinion on Covid, it seems that these sorts of oddities in the refereeing process and online are not unusual. It can not be healthy for the checks and balances of science to be distorted by groupthink in support of a single narrative in this way. What is especially disturbing is the distortion caused by rapid publication of results supporting a standard narrative, with obfuscatory hurdles and lengthy delays being put in the way of studies suggesting a problem with that narrative. Policy based on science produced in this way is likely to be suboptimal.

On point 4: When I was examining the Flaxman paper it was possible to fully replicate what the authors had done. They had provided code that was relatively straightforward to use and check, and documented it adequately, and the description of the model used was presented precisely and concisely. Replicating Report 41 was more difficult. The code was only available via a series of interlocking software packages which made it difficult to follow many aspects of how the modelling was implemented. I had to resort to re-implementing the model from scratch on the basis of the model description given in the supplementary material of the report. That material was exemplary - the reports authors had written down exactly what they had done and replication was possible, albeit with considerable effort. However, the model is only part of the story. Data is the other part. The data to which the models were fitted was made available for report 41, but the data used to determine many of the model parameters was not. These data came from the NHS CHESS database and were confidential. There was no way of checking or replicating the analysis of these data. Journal reviewers will not have attempted anything like the level of scrutiny to which Ernst Wit and I subjected report 41. The same is true for the Flaxman paper, as can be seen by reading that paper's reviews, which are public. While replication of report 41 and the Flaxman paper was possible, it did not appear to be feasible for either the Warwick or LSHTM models that formed the basis for the Omicron advice. Neither provided code, and the LSHTM model description simply cited earlier work which in turn cited other work in a way that made reconstruction of the exact model structure impossible (at least for me). The Warwick model description was clearer, but again did not provide all the details that would be required for full replication, even if all the required data had been made available, which they were not. Clearly time pressures play some part in this, but only a part. The models used were not produced out of thin air at short notice, they were existing models, tweaked to make Omicron predictions. If such models are to be used to produce advice which may result in major

economic disruption, it would be prudent for the modelling teams to be resourced to produce results that are fully and readily reproducible. Proper scrutiny is an essential part of ensuring that results can be relied upon, but is only possible if the results are reproducible. In the pandemic situation that has to mean *rapidly reproducible in practice*, not merely reproducible in theory given weeks or months of effort.

I should add that the tendency for studies to be 'reproducible in theory' but not reproducible in practice is by no means limited to epidemiological modelling. For example the Wang et al. study on global excess mortality, referred to in point 6 and published in the Lancet, states that the data they use are available at the GHDE website, but the link they provide leads to model results, not the data used as inputs. The paper does provide a working link to the code used in the analysis, but this comes without documentation explaining how it is to be used or how to obtain the required data files or their format. There is therefore no practical way that a referee would be able to replicate the results as part of peer review. The descriptions of the statistical procedures provided in the supplementary material for the paper are also not sufficient to be sure exactly what has been done (at least not to me, and I have written textbooks on the types of model used). This paper, which I think is generally a good study, is not exceptional in this regard.

On point 5, one obvious failure to appreciate basic statistical facts is the persistent belief that it is reasonable to use reported cases as if they could provide a reliable index of prevalence. A less far reaching example of sufficient statistical expertise sometimes lacking is provided by my own work on inferring incidence directly from death data. This was taken up by at least one group advising government and 'simplified' in a manner that made the analysis demonstrably and seriously incorrect (inferred incidence then declined too slowly, too late). As a further example, when epidemic models are fitted to past data, it often seems that the level of care that has gone into formulating the epidemic models is then abandoned when it comes to formulating the statistical description of how the model's predictions are related to the data being fitted. We pointed out such problems to the authors of Imperial's report 41, but beyond an acknowledgement of receipt we never received any response to the issues raised, and the version eventually published leaves many of the problems uncorrected. This tendency to view statistical best practice as unnecessary (mere pedantry perhaps) is not uncommon. In mid 2020 the Royal Statistical Society was invited to collaborate on a report on the estimation of the R number led by the Royal Society. The ad hoc RSS committee who read the detailed draft report felt that it revealed a number of substantial statistical problems in how R was being estimated and data interpreted, and the committee made detailed constructive comments on what was needed for improvement. It rapidly became clear to us that this level of engagement with the report's substance was not what was required. We felt that what was really being sought was a stamp of approval. Unable to give this, we eventually had to insist on all mention of the RSS being removed from the report. The December SPI-M advice on Omicron[16] is another example of sub-optimal statistical practice in my view. The report gives much greater prominence to the LSHTM results than to those from Warwick (see the plots presented in particular) in a manner that makes the '10% severity' Warwick simulation appear as an implausible outlier. That scenario proved closest to reality in fact. Part of the difficulty was that Warwick manipulated 'severity' in order to also check robustness to vaccine protection assumptions, meaning that while Omicron having 10% of the severity of Delta might be un-realistic, the Warwick scenario labelled '10% severity' was perfectly plausible. But a more systematic problem is this. It seems likely that the advice concentrated on the LSHTM scenarios because they appeared to be less uncertain, and intuitively it seems reasonable to give more weight to the more certain results. But such intuition is only valid if the reported uncertainties reflect real uncertainty. As mentioned above, they can not be expected to do so without attempting to validate the models for prediction. But even on the basis of the available information, it was unreasonable to trust the LSHTM model's reported uncertainties: if you check how the model had estimated vaccine efficacy against hospitalization it is clear that they had neglected levels of uncertainty in this estimate that would have massively widened the uncertainty in their scenarios: the higher precision was illusory.

Most of academic statistics is concerned with using models to learn about the world from data. In situations in which epidemic models are being fitted to data and then used to make decisions of huge

[16]https://www.gov.uk/government/publications/spi-m-o-chairs-statement-on-covid-19-19-december-2021

societal import it would make sense for the statistical component of the work to be done to the highest modern standards. I do not think that this has happened during the Covid epidemic.

Finally, I would like to express an opinion about the comment towards the end of the session that scientists and clinicians should be speaking to the public with one voice and a clear message. I think that this attitude is dangerous. Science is not a democracy, but if it is a dictatorship then the dictator is nature, and nature does not choose to reveal its workings through any particular self selected priesthood of scientists. Understanding what nature is doing is a difficult and messy business, especially when trying to figure things out at pace. The truth is that there is often considerable uncertainty in this process, which is better dealt with by recognising the fact than by suppressing some evidence in favour of a simplistic standard view. What is especially dangerous is if one view is presented as having overwhelming scientific evidence in its favour, when the truth is merely that it is the majority opinion of scientists. This can only undermine public faith in the science for which there *is* overwhelming scientific evidence, such as anthropogenic climate change or the efficacy of vaccination.

(March 2022)