Written Evidence Submitted by
Sir Patrick Vallance, UK Government Chief Scientific Adviser and Head of the Government, Science and Engineering Profession
(C190111)

Historic access of SAGE modelers to data flows in the SARS-CoV-2 response

I was grateful for the opportunity to speak at the committee's evidence session on 16 July on the lessons learned for providing effective scientific advice during this pandemic. I am writing to provide further detail on data flow access limitations at the start of the UK's response. I want to reiterate that the data owners we have worked with to model this response have tried very hard to provide what we needed, and that the purpose of this letter is to address the data flows between the owners and SAGE and its subgroups.

Early UK modelling data requirements

Throughout the response to the Covid-19 outbreak, SPI-M (the pre-existing Scientific Pandemic Influenza Group on Modelling operated by DHSC) have highlighted the need for high quality, detailed and timely data on the occurrence of Covid-19 cases in all settings, along with wider surveillance data, to track progress of the epidemic and to model its trajectory. In the first weeks of the UK's epidemic, it was difficult for SAGE to accurately assess the state and trajectory of the outbreak at that time due to the lack of data.

In the early stages of the outbreak there were three main sources of UK data used by modellers: PHE surveillance data, clinical data from hospital healthcare records from the Covid-19 Clinical Information Network (CO-CIN), and NHS data (for example on hospital-confirmed and suspected Covid-19 cases, bed occupancy and ventilations), provided via sitreps.

At SAGE #6 (11th February) PHE were actioned to ensure there were plans in place to collect the maximum amount of information from returning UK travellers testing positive for Covid-19. At SPI-M (17th February) PHE agreed a protocol of sharing these data with DHSC and SPI-M modellers, in line with pre-existing processes, which resulted in the relevant data sharing agreements being signed by modelling groups.

The first PHE line list data was provided to modellers by PHE (March 6th). Line list data is detailed anonymised individual level data on suspected cases including onset dates and testing data. However, understandably given the high-pace and novel circumstances, in the early stages of the response the collected data often could not provide insights into, for example, whether cases were linked to others that had already been identified. Other PHE data surveillance systems included data from GP consultations and hospital and intensive care surveillance. The latter system, known as CHESS (Covid-19 Hospitalisation in England Surveillance System) was initiated on 15th March and runs through NHS trusts.

CO-CIN was set up by Prof. Calum Semple to inform DHSC at the end of February. The first reports on its data became available to SPI-M as of 11th March once a sufficient number of patients had been enrolled. CO-CIN reports the clinical characteristics of patients admitted to hospital with Covid-19. It was not set up to track the spread of the virus, and so does not cover every hospital across the UK, nor cases of Covid-19 in the community who do not present to hospital. Nevertheless, it has captured a large proportion of UK hospital cases and it has proven extremely useful to both SAGE and SPI-M. It has, for example, been vital in identifying risk factors associated with poor outcomes as well as well as early signs of the prevalence of hospital-acquired infection. The delay between setting up and the data becoming available is understandable, given time was needed for sufficient patients to progress through the stages of their stay in hospital.

Understanding Covid-19 prevalence and incidence in the community, including in those who may not be admitted to hospital or engage with the healthcare system at all, has also been essential. ONS have provided empirical estimates of this through their Covid-19 Infection Survey, the reports of which were made publicly available in May 2020.

NHS data has been important for informing epidemiological modelling analyses. SAGE #15 (13th March) called for the NHS urgently to direct hospitals to input data into the data collection systems such as CHESS. As an indication of the culmination of early data challenges SAGE #15 (13th March) stated, “Owing to a 5-7 day lag in data provision for modelling, SAGE now believes there are more cases in the UK than SAGE previously expected at this point, and we may therefore be further ahead on the epidemic curve”.

SAGEs #14-#16 (10th-16th March) addressed the challenges of accessing NHS data. However, the collection of detailed data required NHS Trusts to manually input these data into systems, which was beyond NHS capacity at that time (SAGE #19, 26th March). The key data source which gave SAGE early insights into pressures on hospital capacity were NHSE’s daily sitreps, which not all SPI-M modelling groups were able to access initially. This was resolved on 25th March with all groups receiving daily updates since.

Moreover, NHSE sitreps were designed for internal management information purposes, and not to inform modelling analysis. This meant that they lacked granular detail, such as how many patients were new Covid-19 admissions, versus potential re-admissions. There has, however, been good collaboration between NHSE and modelling groups to better understand the data streams, and there is work ongoing to review the sitrep design with SPI-M modellers to further enhance the utility of these data for modelling.

Finally, in the early stages of the epidemic, comprehensive data on Covid-19 in care homes were not available to government. In SAGE #25 (14 April) it was made clear that data around cause of deaths in care homes was insufficient for modelling, and that collecting data from these settings needed to be a priority to understand transmission.

Current data access

As we have developed our understanding of Covid-19, we have been able to improve how we track the disease, and therefore build a library of accessible datasets. This has grown alongside academic research on the disease as it emerged globally and in the UK.

A key lesson learnt from this response is that the capacity of organisations and the design of data systems can limit the timely delivery of effective epidemiological analysis. The UK is uniquely positioned to be a leader in this space, with the NHS as a single-payer system which could utilise the longitudinal health and care records of over 65m people for near real-time analysis in the public interest. Whilst there are many ongoing initiatives across organisations such as NHSX and HDRUK to make this happen, this experience highlights the remaining challenges of variable data quality, fragmentation and secure access to clinical data at scale.

I hope this information is helpful in clarifying what data was available to SAGE in the initial stages of modelling SARS-CoV-2, and in explaining where and why there were initial limitations to access.

Yours sincerely,

Patrick Vallance

(24 July 2020)