The Challenges of Data in Future Pandemics

Nigel Shadbolt[1,2], Alys Brett[3,6], Min Chen[4,6], Glenn Marion[5,6], Iain J. McKendrick[5,6], Jasmina Panovska-Griffiths[7,8], Lorenzo Pellis[9,10], Richard Reeve[11,6], and Ben Swallow[12,6]

[1]Department of Computer Science, University of Oxford, UK
[2]The Open Data Institute, London, UK
[3]UKAEA Software Engineering Group
[4]Department of Engineering Science, University of Oxford, UK
[5]Biomathematics and Statistics Scotland, Edinburgh, UK
[6]Scottish COVID-19 Response Consortium
[7]The Big Data Institute, University of Oxford, UK
[8]The Wolfson Centre for Mathematical Biology and The Queen's College, University of Oxford, UK
[9]Department of Mathematics, University of Manchester, UK
[10]The Alan Turing Institute, London, UK
[11]Institute of Biodiversity Animal Health & Comparative Medicine, University of Glasgow, UK
[12]School of Mathematics and Statistics, University of Glasgow, UK

July 2021

Abstract

The use of data has been essential throughout the unfolding COVID-19 pandemic. We have needed it to populate our models, inform our understanding, and shape our responses to the disease. However, data has not always been easy to find and access, it has varied in quality and coverage, been difficult to reuse or repurpose. This paper reviews these and other challenges and recommends steps to develop a data ecosystem better able to deal with future pandemics.

Introduction

It is a commonplace that we live in a world increasingly dependent on and defined by access to data. In the context of infectious disease control every facet of the understanding we are seeking to achieve is underpinned by data. Other papers in this volume demonstrate that such data varies considerably in quality and coverage. In many critical areas the data may not exist or is in a form that makes it difficult to use. In this paper we explore these issues and relate

the data challenges to particular aspects of the COVID-19 experience, as well as looking at the prospects and recommendations for a more effective data ecosystem to deal with the challenges of this and other infectious diseases now and in the future.

Work in estimation and modelling within the infectious diseases community identifies key data required for current models. Models require data to generate estimates but can sometimes also help identify where and what type of additional data might be required. In particular, issues around the resolution and granularity of data arise. Sometimes these requirements are specific to the questions being raised and the models being used. The data we have available may have been collected for a variety of reasons, some may never have been intended for use in pandemic modelling. The data will be of variable quality and coverage. The modellers themselves may choose to further abstract, simplify or leave out available data. No set of data can be expected to meet every contingency. We can, however, attempt to learn from recent experience to deal with the data challenges that confront the infectious diseases community.

One useful perspective is to try to describe and ultimately anticipate the types of questions posed to our models. From the abstract to the very applied, these questions will be one important driver of the data we generate, collect and curate. Questions posed during the COVID-19 pandemic have included the real-time estimate of the latest reproduction number R and whether different variants have different transmissibility or severity values, general advice on whether it is worth closing airports, schools or universities to reduce transmission and how the effect of these non-pharmacological interventions might vary from location to location or through time. These issues and the data challenges they present are discussed in more detail in section 1, with a view to anticipating similar challenges in future infectious disease outbreaks at whatever scale.

Data, models and results are themselves data objects to be managed. Some data is collected from the physical world and our various digital interactions. Other data is generated *in silico* by our models and used in subsequent analysis. How these different data should be represented, managed and maintained is discussed in section 2 which argues for the importance of FAIR principles in meeting future pandemic data challenges.

How these principles might be realised in data management platforms and lifecycles is discussed in sections 3 and 4. There are particular areas that have worked well, such as the use of open-source repositories and registries, and the development of safe access to large amounts of sensitive linked patient data. The broader ambition to develop consistent data lifecycles and pipelines with which to analyse infectious diseases data is still work in progress.

In section 5 we take a data skills perspective and consider how we ensure that we have the human capital to tackle the data challenges of the next pandemic.

Finally, many of the challenges around pandemic data relate to stewarding data, both in terms of institutional and policy responses. Sections 6 and 7 outline the challenges and potential responses to these issues.

Data plays a crucial role in combating infectious diseases. This is not a new insight(Heesterbeek, Anderson, Andreasen, et al., 2015), but the importance of such a role will only increase in the future and it is essential that we attempt to construct effective, efficient and equitable data ecosystems. Figure 1 attempts to illustrates the major components and dependencies of such data ecosystems, and which we examine in more detail in Sections 1 – 7.

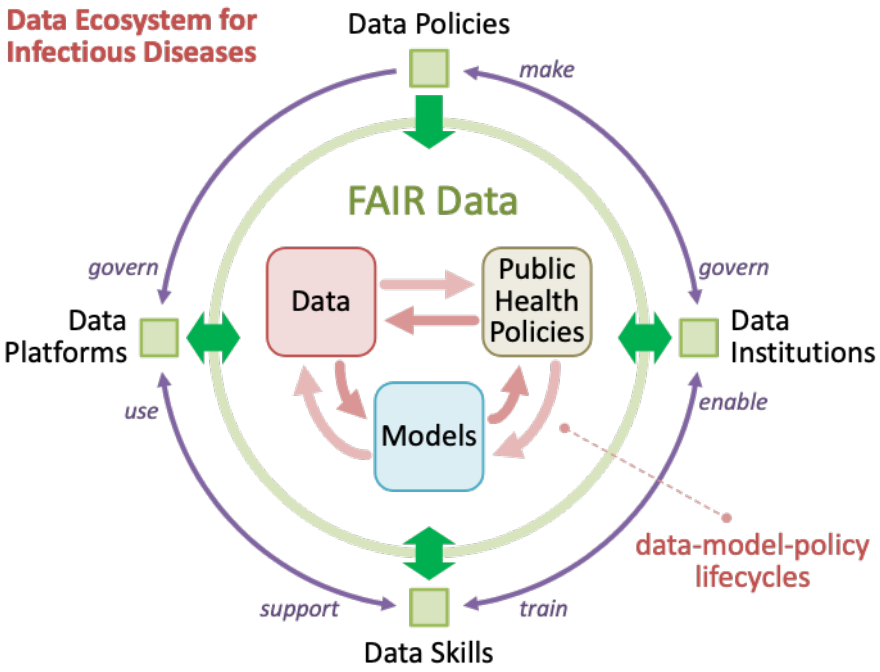

Figure 1: Key components and dependencies in data ecosystems for combating infectious diseases. The FAIR data principles (Section 2) should underpin any data ecosystem. Data-model-policy lifecycles (Section 4) are at the heart of a data ecosystem, while data platforms (Section 3), data skills (Section 5), and data institutions (Section 6) represent the physical, personal, and organizational entities of data ecosystems. Data policies (Section 7), which are formulated by data institutions, should embody the FAIR data principles and govern the data ecosystems.

1 Data and Models

Data is crucial when modelling pandemics. Data is needed to parametrise and validate models that condition projections and scenario analysis of future epidemic trajectories. These trajectories are used to inform public health policy decisions.

1.1 Data Availability

The emergence of collated data sources assembled by the research community was seen early in the COVID-19 pandemic. Initially data was sparse, early sources included data scraped from news outlets, press briefings updating daily case counts, social media, crowd-sourced and open data (Wikipedia, n.d.; Imai et al., 2020; Read, Bridgen, Cummings, Ho, & Jewell, 2020; Sun, Chen, & Viboud, 2020; Wu., Leung, & Leung, 2020). As the epidemic developed into a pandemic, data collation became more systematic, resulting in extending existing repositories or developing new ones, e.g. the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (Dong, Du, & Gardner, 2020) and Our World in Data (Our World Data, n.d.). Systematic reviews of data from case studies also started appearing (Oxford COVID-19 Evidence Service, n.d.), with the volume of freely accessible data increasing, including development of public data dashboards by individual governments.

An important source of data were volunteer-based initiatives to curate high quality data, freely available and including associated methods and code, e.g. the Google GitHub repository of COVID-19 open data (GitHub GoogleCloudPlatform, n.d.). Despite the practical issues with inconsistencies and lack of centralised oversight in crowd-sourced efforts, these repositories offer

a glimpse of what is possible by recruiting community participation on large platforms. The large platforms have also been able to provide data directly generated from interaction with their users. Facebook providing data on movement, symptoms, vaccination status, and people staying in place (Facebook, n.d.), Google providing a wide range of mobility data (Google, n.d.), whilst Apple provided further mobility insights around the mode of travel (Apple, n.d.).

Modellers also had access to data via the involvement of national agencies, though it should be noted that due to privacy concerns not all data collected by such agencies were widely shared. In England and Wales, this role was played by the ONS (Office for National Statistics, n.d.-b)and the COVID-19 infection survey (Office for National Statistics, n.d.-a) and in Scotland by Public Health Scotland (Public Health Scotland, n.d.), which provide important data on general levels of the virus within the UK population. In addition to the statutory bodies, a number of research council funded initiatives at leading university labs undertook a range of studies that continue to furnish insights into population level transmission dynamics (e.g. in the UK REACT (Imperial College London, n.d.)) and the impact of Non-Pharmaceutical Interventions (NPIs), adherence to Test, Trace and Isolate guidance and other measures as the epidemic developed (University College London, n.d.; Smith et al., 2021). Other innovative data sources have included the use of wastewater COVID-19 RNA sampling, sampled from sewage catchments at a variety of spatial scales, and thus involving a wide range of public and private bodies (Wade et al., 2021).

The UK has seen the use of organisations such IPSOS Mori (Riley et al., 2020) and YouGov (YouGov, n.d.) with online survey methods and data analytics that have augmented the health and research sectors' ability to capture near real-time data on a range of pandemic issues. In Germany the COSMO project surveyed differences in risk perception, knowledge and protective behaviour regarding COVID-19 (Rattay et al., 2021), the Netherlands produced analysis on changing behaviours of work and travel, building on already established census-like questions (de Haas, Faber, & Hamersma, 2020).

Another type of data has increasingly been incorporated into epidemiological research: happenstance data - "not originally collected with a particular research or policy question in mind but... created through the normal course of events in our digital lives, and our interactions with digital systems and services." (The DELVE Initiative, 2020). These data can help us understand everything from economic activity to population movements. However, their use is limited by a variety of challenges around availability, anonymisation, interoperability and consistency, as well as the fact that they often only offer a proxy for contact patterns relevant to infection spread. One such example, mobility data, has been used to understand clustering within society and deriving modelling assumptions when evaluating contact tracing. It has also proved important in determining the efficacy of NPIs such as various forms of lockdown (Zhou et al., 2020). Another key mobility data source was flight data – e.g. data on the huge migration out of Wuhan for the Chinese New Year in advance of their lockdown proved to be important (Song, Zang, Ding, et al., 2020).

Recommendation: the continuation of open data publication and dashboards for infectious diseases as a part of civilian health infrastructures modelled on those created during the COVID-19 pandemic and in a form that is useable in aggregate fashion by all or in more detail under licence to approved researchers. To include happenstance data that has proven valuable, such as aggregate mobility data, and survey data. Continued support for established registries and repositories of infectious disease data.

1.2 Data and Model Parameterisation

One aspect of mathematical modelling that is data reliant is the fitting process. In the simplest case of pure exponential growth or SEIR population-level models this means matching the temporal trajectory of the number of infectious people to data on reported infections and/or matching the mortality over time in the model to the data. While matching such aggregated data to model estimates will give us an informed picture, there is a large advantage in using stratified data to validate the model (Balabdaoui & Mohr, 2020). For example, assessment of the effectiveness of vaccination strategies based on age priority *a priori* requires the data to be age-stratified. In this case it is important to be able to match hospitalisations and deaths in cohorts prioritised for immunisation. If this is not achieved, the overall agreement of the model with the data may be as good, but the impact of vaccination may be poorly predicted.

The reality of modelling for policy decision making is that policy stakeholders typically commission research groups to assess specific policy questions. Under "normal" research timescales (defined here as outside of a global pandemic) this is usually achieved via a tender application. As part of the collaboration, data that would not usually be accessible to researchers would be made available. During the pandemic, this process has continued but at a much larger scale and faster pace.

Around the world modelling groups have had to respond to urgent policy questions from their governments. In the UK, the SPI-M modelling group has been responding to numerous commissions from SAGE on a number of policy questions posed by the UK Government, beyond initial work that informed the need for and timing of lockdown (Ferguson et al., 2020). These included exploring different scenarios for reopening society after the first epidemic wave in conjunction with Test-Trace-Isolate (TTI) strategies (Thompson et al., 2020; Panovska-Griffiths et al., 2020), evaluating emerging questions such as the transmissibility and severity of the B.1.1.7 variant (Davies et al., 2021; Challen et al., 2021), or formulating the roadmap of reopening society as vaccination roll out continues (Whittles et al., 2021; Keeling, Moore, Dyson, Tildesley, & Hill, 2021).

Data is key to correct modelling and analysis in all these scenarios (The Royal Society, 2020; Bowman, Silk, Dalrymple, & Woods, 2020). For example in the UK SPI-M has had access to a large portfolio of data, in addition to what is publicly available. Different modelling groups, both within SPI-M and outside, have used different data, ranging from mortality, hospitalisations and serology data, to data on reported infections and from the ZOE COVID Symptom Study or REACT, with most models combining more than one data stream.

It is important to note that the choice of data remains at the core of validating and using mathematical models; it is indeed sometimes difficult to untangle when and where the data analysis stops and the modelling starts. We discuss some of these challenges in our discussion of data lifecycles and data skills. In all cases the existence of richly described data under the FAIR principles, which we discuss later, needs to be a new normal.

A clear lesson from the SARS-CoV-2 pandemic is that modelling was undertaken even with limited data. In fact, the earliest data, even if scarce, is often invaluable for key estimates especially of hard-to-measure quantities (e.g. incubation period (Overton et al., 2020), likely number of unknown infectious cases and probability of case ascertainment(Omori, Mizumoto, & Nishiura, 2020) etc.), thanks to the geographical circumscription of an initial pandemic phase: given infection is rarely observed, detailed travel history to the source location can be used to bracket the exposure window. Key estimates, for example of epidemic doubling times, are then revised as more data becomes available (Pellis et al., 2021). Access to relevant dispersal data becomes essential.

We also know that different kinds of models require different amounts of data. SEIR and branching models require less data than structured Agent Based Models (ABMs). More complex (and hence potentially more realistic) ABM models can capture more fine-grained behaviour. It is important though to strike a balance between model details and parsimony. Achieving this is not always easy, and often, in an emergency, modelling groups will use readily available models (e.g. established models for influenza spread (Coburn, Wagner, & Blower, 2009)), supplementing what cannot be informed from currently available data with assumptions. For example, in the current pandemic, early studies assumed generation times typical of SARS-1 or MERS30, and scaled versions of pre-pandemic contact surveys (Mossong, 2008; Klepac, 2020), were used to approximate contact patterns during lockdown, before the CoMix (CoMix study, n.d.) study was set up. The development of complex ABMs that make use of large and complex data, including better contact and behavioural data, has been one branch of modelling that has developed rapidly during this pandemic (Kerr et al., 2021; Hinch et al., 2021). Building a publicly available suite of commonly used models with variable degrees of complexity would be extremely valuable, and there are various attempts in this direction (Epirecipes, n.d.).

As more data becomes available, models can be expanded or revised, if the modellers are aware of its existence. This work, together with new policy questions, might indicate the need for new data. However, some aspects of the model will never be properly informed by data: for example, the uncertainty regarding the impact of NPIs or unknown future changes in policy mean models often need to make strong assumptions, and sensitivity analysis to assumptions and exploration of wide spectrum of scenarios remains key for scientific robustness and policy advising.

Though always desirable, more data also gives rise to new challenges, such as inconsistencies between different data streams, or predictions for the same measure from models fitted to data from different sources or collected differently – e.g., in the UK, REACT, ONS or COVID-19 dashboard – might differ. Often the reason might simply be that two data sources might contain apparently similar but actually different data (e.g. in the UK, ONS reports all deaths weekly (Office for National Statistics, n.d.-c) while the COVID-19 dashboard (Data Gov UK, n.d.) reports deaths within 28 days of a positive test). It is therefore imperative that model assumptions and caveats are clearly stated, and details about the data used, alongside with rich metadata, explicitly stated. However, data will always evolve as a pandemic progresses, so the practical details of this inter-dependence of models and data will always need to be dealt with in the contingent situation.

Recommendation: Data needs to be re-presented and re-structured in many different ways as questions come to the fore. Data should be captured with a broad and rich set of attributes, made available under the FAIR principles, so that at a minimum the existence of particular data assets is a matter of public record. In all reported work details of the data used, its limits as well as its utility, and model dependencies need to be made explicit. The lifecycle of model evolution, data and model dependency needs to be captured.

1.3 Data Types

The COVID-19 pandemic has highlighted the range and type of data required to deal with a severe infectious disease at scale and pace. It has highlighted a clear need for stratified summary statistics. Ideally as fine grained as possible, describing the who, what, when and where of the disease trajectory as well as heterogeneities in the at-risk population. In the face of an emerging pandemic, sharing timely and reliable data has been challenging. For example, in many countries rapidly changing levels of infection meant the ability to obtain reliable and

representative test results turned out to be extremely challenging (The Lancet Respiratory Medicine, 2020).

One objective outcome that became more reliably captured in data relates to hospitalizations. The number of admissions, deaths in particular health boards, authorities and regions of those admitted with COVID-19, information on the age structure, ethnicity, socio-economic status, profession, underlying health conditions have turned out to be essential data in better understanding the nature of the disease (Navaratnam, Gray, Day, & Wendon, 2021). The duration of time spent in hospital and in what context – ICU or ITU and associated recovery rates – have proven important in managing resources and outcomes. The extent of illness in the wider community and any COVID-19 deaths that were not associated with hospitalization have also been important data to attempt to qualify our more certain hospital data. In the UK, both REACT and ONS surveys provided key data. Whether this routine health data surveillance continues is an important policy choice going forward.

The results from testing health workers, both positive and negative, became an important data resource. There are clear challenges around reliable and effective test data, both test-based (and serological) surveys necessarily reporting estimates of sensitivity and specificity for the tests used. There has been an increasing awareness of the importance, when testing, of pathogen genotyping information where available. The role of crowd sourced self-assessment tools has also yielded interesting data - symptom tracking (e.g https://covid.joinzoe.com) and self-reporting of, for example, self-isolation.

A particular challenge is the move from summary data to detailed data on individuals – the more that is routinely collected the richer the inferences we may be able to make (((Marion et al., 2021) and (Swallow et al., 2021) in this special issue). Times of: admission to hospital, testing (multiple tests incl. previous test negative score if available), entry and exit from ICU and time of discharge or death, proxy info of time of infection for early cases (e.g. in Europe those returned home from Wuhan or northern Italy); and characteristics of individuals e.g. age, gender, underlying health conditions, ethnicity, settings etc. if available: informati about timing of treatment, genetic information etc. However, the more detailed the data, the more pressing the issues likely to arise from privacy legislation and the greater the need for appropriate metadata.

The experience of this pandemic has further highlighted the need in future for fundamental demographic / denominator data – at as high a resolution as possible. These include

1. population distribution at high spatial resolution broken down by age, gender, ethnicity, social deprivation indices etc.;

2. health care demographics: number of: beds in ICU/ITU and number of health care workers; number, distribution, occupancy and staffing rates of care homes; current (and planned) location and capacity of testing centres; contact tracing capacity e.g. public health workers;

3. social structure e.g. household structure, school catchments, workplace size distribution at high spatial resolution;

4. mixing/contact info: e.g. patterns of mixing or relevant data such as GPS tracking or output from behavioural science simulations (e.g. urban analytics) that enables estimation of mixing under different control scenarios within and between individuals with different characteristics (e.g. age, gender ...) and different locations (e.g. areas within a city, cities, towns and rural areas). Ideally these would be combined to give mixing

between individuals at location x_1 with characteristics c_1 and those at location x_2 with characteristics c_2.

5. Settings with different risks of infecting others and being infected, e.g. workplace, university, asylum centres, prisons, schools, retail, hospitality etc.

A further category of data that needs to be captured relates to data about interventions and public health measures:

1. nature and dates of health messaging and awareness of the epidemic;
2. nature and dates of public health NPIs e.g. lockdowns and other restrictions imposed or advised (The Health Foundation, n.d.; Blavatnik School of Government, Oxford, n.d.);
3. information on vaccination over time broken down by relevant factors e.g. age, gender, location, occupation, socio-economic status.

The need for data linked to time (e.g. dates) as information is lost by integrating over time scales that are large relative to the outbreak dynamics. Data should also be stratified as much as possible since epidemiologically useful/relevant information is lost when summarised over a heterogeneous population.

An example of data that was scarcely available at all at the beginning of the current pandemic, but has since became crucial, is genetic sequencing data (Nextstrain, n.d.), a key data source with the emergence of viral variants (Public Health England, n.d.), to monitor how their prevalence varies across communities and across time, both within host and between human hosts.

The pandemic has highlighted data gaps in various countries, gaps that have proven critical in combating its effects. Care homes have been badly hit around the world, but data on them has been, and still is, extremely scarce (e.g. in the UK in the first half of 2020, data from care homes was limited to the number of beds in the home and the presence or absence of an outbreak, not even how many cases or deaths). Much more was known about schools, because influenza is concentrated there and previous surveillance measures were in place. The challenge is to develop reliable data collection procedures in various societal settings. Prisons are another case where data is very scarce. Covid in common with many infectious diseases thrives in enclosed communities.

The DELVE report recommends, and the authors of this paper also endorse, a step-change in the use of mobility data for public policy. One means of achieving this would be to have statutory bodies like the ONS act as the trusted agency to convert happenstance data into high-frequency population mobility statistics. An alternative could be to establish a Data Institution to manage such data (see Section 6). The ambition would be to produce, from mobile phone operators, daily views of population mobility between geographic regions, aggregated from origin to destination counts. It would also be valuable to modellers if such data were available for retrospective analysis and not deleted after some arbitrary period, but this must necessarily be balanced against privacy concerns associated with the data and its retention.

Recommendation: Large-scale infectious disease outbreaks require data assets to be available or capable of being captured reliably at an appropriate scale. This will include; (a) stratified summary statistics on the epidemic by age gender, location etc., (b) the same data on individuals, pathogens and variants, tests, treatments and vaccines, (c) demographic or denominator population data at as high a resolution as possible including mixing rates between different subpopulations and details of populations targeted by testing and other interventions, (d) data

on interventions and public health measures from tests to vaccines, NPIs to communications. Care should be taken when updating datasets with enhanced information to avoid needlessly altering datatypes and structures, however, to avoid difficulties for those working with existing resources.

2 FAIR Data

The quest for effective scientific data management and stewardship has a long history, restated in 2016 under the FAIR (Findability, Accessibility, Interoperability and Reusability) principles (Wilkinson, Dumontier, Aalbersberg, & Appleton, 2016). These sought to lay out a set of guidelines to maximise the utility and potential reuse of data. A particular emphasis was on standards that would support machine-based discovery and processing of data. These guidelines rapidly gathered endorsement, for example at the 2016 G20 summit (G20 Leaders, 2016) and figured in a recent statement from the G7 Science Academies (Science Academies of the G7, 2021). They are adopted by international organisations such as CODATA (CODATA, n.d.) and the Research Data Alliance (Research Data Alliance, n.d.) , organisations committed to building research data ecosystems that can solve cross domain challenges.

Figure 1 illustrates the fundamental role we see for the FAIR data principles in developing, managing, and executing data-model-policy life-cycles (see Section 4). Data policies (see Section 7) should adhere to FAIR data principles and be realised in the implementation of data platforms (see Section 3). Data institutions (see Section 6), including public health bodies, should be informed and guided by the FAIR principles when formulating data policies or participating in the data-model-policy lifecycles. Understanding the FAIR principles should be a key skill for all those engaging with the data-model-policy lifecycle (see Section 4).

However, the reality is that, whilst there is some increase in adoption of the principles, many epidemiological modellers remain unaware of them and, even where they are recognised, any concrete or consistent implementation has lagged far behind. In the following subsections (2.1-2.4) we present a finer grained analysis of the challenges presented by the FAIR principles, relevant experience during the COVID-19 pandemic, and recommendations to help realise them within the infectious diseases community. We also highlight areas (2.5 through 2.7) where we believe it is important to supplement the FAIR principles.

2.1 Findable Data

The principle of Findability requires that data are assigned globally unique and persistent identifiers (Meadows, Haak, & Brown, 2019) and are described with rich metadata (DataCite, n.d.-b). It also requires that data are registered or indexed in a searchable resource.

Several challenges surround this simple principle of Findability. As noted in section 1.1, many qualitatively different types of data are relevant to pandemic preparedness, modelling and control. These include health data (sensitive at a fine spatial or demographic scale and held in the first instance by individual health care providers), the underlying baseline demographic data (also sensitive at fine scale, held by government statisticians), movement data, both individual daily movements and long-distance travel (often commercially confidential and held by companies as part of their routine service provision), and possibly including datasets as varied as weather records and predictions (held by meteorological organisations) and social care data (often held by private sector organisations that run care homes, and if aggregated at all often only at a sub-national level). There is a need to promote and embed FAIR principles in the data

management practices of these various holders of data. Even if they are variously inclined to make the data available, the knowledge of the data's existence is critical. Persistent identifiers and open metadata are a fundamental feature of any data infrastructure (Clark, 2021).

Over the past decade numerous open data projects around the world[1] have sought to make available access to key data sets from government departments and public services. These efforts have extended to universities, research labs and funding agencies and have been informed by the FAIR principles. During the pandemic many governments have set up coronavirus extensions to these efforts (see for example `https://coronavirus.data.gov.uk` in the United Kingdom).

Very different findability standards apply to these different sites, however. Many sites set up during the pandemic have poorly described data, with little, if any metadata. An example of good practice is Open Data Scotland (`https://www.opendata.nhs.scot`) with detailed metadata on every individual dataset, containing both versioned records of the dataset metadata, metadata on the individual columns in every dataset, and even versioned higher-level data on different column entries, with associated metadata about these new meta-datasets.

However, sensitive data is another matter, none of these sites provide access to metadata on non-Open Data, and there is certainly no searchability and persistent identifiers that make sensitive data findable if the researcher is not already intimately familiar with it. This strongly limits the utility of such data.

For sensitive data, metadata are often treated as if they were as sensitive as the data they describe. It is legitimate to consider the sensitivity of the metadata itself, but it's not clear whether data that is so sensitive that we're not allowed to know it exists should, in fact, be being collected. Modelling activities based on such data will consistently lack transparency. Work is beginning in this area, for instance `https://www.healthdatagateway.org/` in the UK.

Given the enormous range of different and rapidly emerging datasets a single searchable resource is infeasible. Search engines exist that trawl multiple resources with open standards on the metadata. The approach in schema.org is to define vocabulary for providing dataset metadata, alongside (proposed) vocabulary for describing aggregate statistics in ways that can be understood by the major search engines: Google, Microsoft Bing, Yandex and Yahoo!. This is one way to deal with the challenge that a modeller would not know how to search for data they do not know exists. Google's recent Dataset Search capability (Google Dataset Search, n.d.) uses the schema.org format and now anyone who publishes data can make their datasets discoverable in Dataset Search by using the open standard to describe the properties of their dataset on their own web page.

Another, pragmatic, way of discovering new relevant datasets is simply to find which datasets other people are using for related problems. Publicly recording the provenance of model outputs and reports allows identification of the specific data they used. It would support a different kind of findability as well as being important in its own right e.g. flagging of issues once detected. Promoting this kind of openness is a challenge in its own right, and it is discussed separately below.

Recommendation: Policy should require the use of globally unique and persistent identifiers to identify data, including sensitive data, to further open up data or at the very least publish relevant metadata. Metadata needs to be released under open licenses and schemas so that it can be archived, preserved and transferred to other searchable resources if necessary, or else used in initiatives such as schema.org and be part of the normal reporting and publication process for

[1] See for example `https://data.gov` in the US, `https://www.data.gouv.fr` in France, `http://data.gov.uk` in the UK and the EU open data site `https://data.europa.eu`

models and their data.

2.2 Accessible Data

The principle of Accessibility requires that data and metadata are retrievable by their identifier using a standardized communications protocol, this may be machine based or involve human intermediaries in cases where full automation is not possible. The challenge is to ensure that data can be obtained by machines and humans upon appropriate authorization.

Non-sensitive data have become much more broadly available during the pandemic. As well as data provided directly by government, there have been several major initiatives to curate high quality data on large information technology platforms.

We noted in section 1.1, the Google GitHub repository of COVID-19 open data (GitHub Google-CloudPlatform, n.d.) which provides an example where the process of depositing, publishing and sharing data is mediated through significant support from a community of data wranglers. These data are all available through simple protocols such as https.

Access to sensitive data is more complex, but approaches have been developed during the pandemic to make them more accessible. OpenSAFELY (OpenSAFELY, n.d.)is a new kind of secure analytics platform for electronic health records in the NHS, which has been created to deliver urgent results during the emergency. It currently delivers analyses across more than 55 million patients' full pseudonymised primary care NHS records. We discuss it in more detail in 3.1.

The technical Accessibility requirement is usually met by the use of 'click on the link' protocols such as https. However, where restricted access may make fully automated access impossible a contact protocol in the associated metadata is a requirement. The technical Accessibility protocol should be free and universally implementable whilst also allowing for authentication and authorisation. Metadata should always remain accessible even when the primary data is access restricted, no longer available or deprecated.

Recommendation: Standards should be agreed so that researchers needing access to new restricted data sources can do so without high barriers of entry in an emergency.

2.3 Interoperable Data

Interoperability is ensured through the use of broadly applicable knowledge representation formats such as JavaScript Object Notation - JSON (Marrs, 2017). These provide an open standard file format and data interchange format.

Interoperability is also supported through the use of widely adopted controlled vocabularies, such as SNOMED CT, UniProt and W3C DCAT (W3C, n.d.-a) to describe the metadata itself. EMBL-EBI maintains a virus and general infectious diseases ontology as well as a more specific COVID-19 extension using the OWL Description Logic as the knowledge representation language (eMBL-EBI, n.d.). Despite the ontology lookup service mapping into other widely used controlled vocabularies such as UniProt(UniProt Org, n.d.), it does not appear to have been widely used.

There is little evidence that outside of particular areas such as medical records data the epidemiological modelling world was unified in adopting agreed, widely adopted controlled vocabularies.

As to other forms of data that are available for incorporation into models such as happenstance

data, they rarely use standard ontologies, and may not even be in standard formats.

Recommendation: The uptake of data in future pandemics requires the development of pipelines for converting data quickly and easily, while respecting its provenance, into more standards-compliant forms so that it can then be provided in a FAIR manner.

2.4 Reusable Data

Reusability requires rich descriptions of data along with a plurality of relevant and accurate attributes. Each record should contain a set of properties that meet a level of descriptive adequacy to help ensure reusability – see for example DataCite's mandatory terms for metadata (DataCite, n.d.-a).

Reusability also requires that data is supplied with a clear and accessible data usage licence, detailed provenance and meets domain-relevant community standards. Provenance information allows for a more detailed understanding of the origination, and subsequent history of any modifications of the data. Standards exist for this (W3C PROV) (W3C, n.d.-b). We discuss this further in section 2.6, and also provide an example of a platform for monitoring provenance (the FAIR Data Pipeline, section 3.2).

Recommendation: Data resources should be richly described with a plurality of accurate and relevant attributes.

2.5 Linked Data

Datasets containing information on specific aspects of people's private lives is inherently sensitive, and the ability to link these aspects together presents challenges. There is an understandable reluctance in allowing a government to build a full picture of someone from where they live to where they work, from when they travel to where they sleep, from who they call to who they meet, from the specific viral strain they are infected with to which other people have been infected with closely related strains.

However, being able to make inferences and detect patterns across these linked sets of individual data can help enormously in controlling the spread of an infectious disease. In Scotland the Community Health Index Number (CHI number) uniquely identifies a patient, and since 2013 it has been possible to use it for social care and other non-health service bodies. It facilitates powerful studies (SLS-DSU, n.d.) although for some it raise concerns around privacy and proportionality.

Is it possible even in principle to preserve privacy while connecting all of these datasets together? If not, then do we compromise our right to privacy or our ability to accelerate the control of infectious diseases in an emergency? This is not purely a technical concern – approaches such as OpenSAFELY, discussed later, may allow linkage to be carried out in a secure manner, but if the data exists to link records, who is to say that it will only be used during a pandemic and in a controlled way?

If linkage is possible in a suitably privacy-preserving manner, then how do we link datasets in practice that are held in different TREs under different restrictions? One solution would be to hold all of the data in a single TRE, but most countries could not consider sharing such data with others for legal reasons and, even if it were possible at a more local (e.g. national) scale, the sheer variety of data and increased risk of misuse may make it practically impossible. Another is to federate TREs to allow analysis of data that sits at rest in multiple locations

(HDRUK, 2021).

Recommendations: Policy decisions need to be made around proportionate linkage both of data in purpose-limited ways, involving agile methods to determine context-specific consent from policy makers and and also critically from engaged citizens. Further work needs to be carried out on different approaches to federation and aggregation of linked data assets between and within jurisdictions and providers.

2.6 Provenanced Data

Provenance issues arise when considering the FAIR principles, but they are equally important when we consider the pipeline from data to policy. We have already noted that data availability indirectly impacts model structure, but data directly drives model outputs, and model outputs inform reports summarising the current state of a pandemic or predicting its future course under different control scenarios. These modelling reports then affect policy proposals and advice, and these in turn inform policy. However, very few of these dependencies are exposed or represented in a way that allows for inspection. Even when reproducibility concerns are highlighted, what data modelling code uses and what outputs it generates are often opaque to anyone but an expert, often only to one in the research group running the model.

We have noted that a common standard exists for tracing provenance (W3C PROV), and this could be used for tracing data use in modelling and analysis. It will be a major challenge to enable, much less enforce, adoption of standards and practice that will allow us to trace the path from data through code to policy adoption (Harris et al., 2018), but we believe it is critical to allow us to understand how decisions are informed, and what confidence we can therefore have in them. We discuss one tool that enables such tracing (the FAIR Data Pipeline) in section 3.2 below.

Recommendation: Develop provenance and accountability traces for the full life cycle of acquisition, through model use to policy advice.

2.7 Quality Data

A significant failing of much data collected, and particularly that available in the early stages of a disease breakout is its quality (Wynants et al., 2020). This also applies to data not ordinarily collected or used for analysis. More importantly, metadata describing the limitations of data is often lacking or not easily interpreted and/or machine readable, making the data difficult to analyse correctly. Missing data, or data removed because of privacy concerns, is often hard to identify and handle correctly. Furthermore, issues that are identified with datasets, either before or (even harder) after use, are ordinarily recorded (if at all) separately from the data and are therefore almost impossible to trace and use to assess confidence in results arising from their use.

When assessing data quality as it impacts on modelling activities, the approach proposed within the Scottish COVID-19 Response Consortium was to explore these key attributes:

- the degree to which a dataset was well defined and documented, and had good metadata;
- the 'intrinsic quality' of the data, focussing on concepts such as credibility and completeness;
- uniqueness, given that some information was available from very few sources;

- relevance or applicability, as a case-specific assessment of how well aligned was the dataset with the needs of the model; this is not a property of the dataset, although it depends on these.

The aim was to combine the assessed criteria to give an overall, ordered categorical, score for 'fitness for purpose', and thus identify datasets which should not be used to support modelling, and to prioritise which other data resources merited most urgent attention. Other criteria might reasonably be considered, but it was felt that these were the most important in practice. FAIR principles aim to facilitate reuse, but their use also supports decision making about many of these criteria. By assessing a dataset as green (low priority for further investigation), amber (mildly problematic, but not a priority), red (high priority for further investigation of alternative resources), or black (not appropriate for use), this approach aimed to support ongoing continuous improvement in available data resources, while itself generating useful metadata to associate with datasets.

Recommendation: Mechanisms are needed to record data and code quality issues that, together with provenance traces, will allow the end user to identify potential problems, even those only identified retrospectively, with both specific model outputs and the derived policy advice. These quality attributes should automatically form part of the associated metadata for any output.

3 Data Platforms

We have discussed some of the challenges around data availability and accessibility, and the extent to which existing data assets have been necessary, though certainly not sufficient during the pandemic. We have detailed a number of responses by governments and other organisations, groups and individuals to make data available. But what of the data platforms that have been devised to deal with the demands and exigencies of large scale and urgent data analysis? And what of pipelines that allow modellers and policy advisors (and the many people who now fulfil both roles) to provide a chain of trust that connects the raw data flowing from these sources to the policy advice that we hope ultimately guides government decision making? We will discuss some approaches that we think provide useful insights into both challenges and recommendations for ways forward.

3.1 Open Secure Analytics

The experience of the pandemic has challenged the normal model of researchers working on intermittently extracted records for restricted numbers of patients via a conventional research data service. This data is commonly months out of date when it becomes available for research. What is needed is a response to the challenges of speed, near real-time situational awareness and scale.

OpenSAFELY is a new kind of secure analytics platform for electronic health records in the NHS, created to deliver urgent results during the global COVID-19 emergency. It currently delivers analyses across more than 55 million patients' full pseudonymised primary care NHS records. It represents a pragmatic, efficient and secure approach that delivered the first analyses showing factors associated with COVID-19 death across 17 million patients, subsequently appearing as a Nature article (Williamson, Walker, Bhaskaran, Bacon, et al., 2020), in just five weeks from project start during the COVID-19 pandemic.

OpenSAFELY uses a new approach for enhanced security and timely access to data, in several

important respects. Firstly, data remains within the secure environments managed by the electronic health record software company. Analysts run large scale computation across near real-time pseudonymised patient records inside the data centre of the electronic health records software company. Secondly, the OpenSAFELY project does not use an 'off the shelf' trusted research environment (TRE). Bespoke software creates a series of increasingly non-disclosive tables to protect patients' privacy, while preserving all the detail in the data for research analysis. Thirdly, the framework generates realistic dummy data. This adds additional privacy protection. Analysts develop all their code collaboratively, and in the open, without ever touching any potentially disclosive patient records. When their code is ready to run, it is sent into the secure environment where it runs against the real data, so analytic code can run against the full raw information. From this secure environment, only the summary results tables and graphs are released.

In addition to these features all analysis code is shared for review and re-use. Analysis code, intermediate codelists, and algorithms are shared openly, by default, in structures that make it possible for subsequent researchers to see how the analysis was done and to efficiently re-use components of the work. The benefits include reproducibility, transparency, trustworthiness, and efficiency. The platform is built in such a way that only shared code can run. In addition, the platform and analytic software is also made openly available for security review, scientific review, and re-use.

Lastly, the paradigm of analysts developing code against synthetic dummy data means that all code is guaranteed to be non-disclosive and shareable, meaning that the platform is able to share in real time a full log of all code ever executed against patients' data, with links out to the GitHub repository and commit ID of each process (albeit that these repositories can, at analysts discretion, remain closed until the analysis results are complete and reported). This builds accountability, by showing patients, professionals and other stakeholders an unambiguous and complete account of all platform activity, allowing any interested party to evaluate whether the datasets are appropriately minimised, and the analyses are proportionate and permitted.

The design is a considered response to the overarching concern of risks associated with data sharing at scale. Particularly as they relate to highly sensitive data. Whilst many models of infection may not require seeding with actual patient data, the understanding of risks, therapeutics and surveillance do. The platform has shown a novel way to tackle this data challenge.

The results obtained have demonstrated the value of this approach. From the first results on factors associated with COVID-19 deaths, to more detailed assessments of patients with conditions such as COPD, asthma (Williamson, Walker, Bhaskaran, et al., 2020) and HIV (Bhaskaran et al., 2020), studies on therapeutic effects of drugs such as hydroxychloroquine (Rentsch et al., 2020)(no benefits or disbenefits observed), and analysis of whether NSAIDs might have an adverse effect on outcomes (Wong, MacKenna, Morton, et al., 2021), valuable insights have been generated.

However, one potential current weakness of OpenSAFELY (along with all existing mechanisms for analysing data in TREs) is that they are constrained by issues of trust between secure environments, restricting analyses to data held wholly within a single store. Currently these issues are only solved by mutual agreements to share individual datasets between stores, a solution that does not scale well.

Recommendation: Develop platform(s) for open secure data analytics such as OpenSAFELY, and investigate mechanisms for data federation to allow individual TREs to control access to their data in a shared data collection, allowing analyses to be carried out seamlessly using data from multiple secure data providers.

3.2 FAIR Data Pipelines

An important concern during the pandemic was the traceability of policy decisions back to primary data and models, and the degree of trust that could be placed in the component parts of this pipeline and in the connections between them. The challenge is therefore to develop software tools to enable traceability of dependencies between analyses that may be used to inform policy, and the models, methods and data on which they are based, and to make such tools attractive and easy for modellers and analysts to adopt.

Such a pipeline would link model outputs to input data and the models used to generate them as well as any assessments made about these, for instance through peer review of data and models. This would create an acyclic directed graph of dependencies that would also allow 'issues' to be flagged and tracked. Likewise software version control systems could be tied into such a pipeline so that issues were raised against previous versions of code bases when commits were flagged as bug fixes, or issues were raised and accepted on software development platforms such as GitHub. Such 'warnings' would apply to any data, model or method of analysis for which issues are identified, and any of their dependencies found in the provenance records at the time of search (not of use).

Such a system should be agnostic on whether or not data are openly available, but all data must be publicly and uniquely identifiable for the result to be FAIR, and the metadata should be publicly available for the pipeline to be as useful as possible, and compliant with the extended FAIR standards detailed in section 2. Standards for unique identification of public artefacts are already well extablished, from Open Researcher and Contributor IDs (ORCID, 2021) for individuals and Research Organization Registry (ROR, 2021) IDs for organisations to Digital Object Identifiers (DOI Foundation, 2021) for documents and datasets. Unique identifiers could also be relatively easily created for datasets held privately, but this would greatly benefit from cooperation with creators and maintainers of data safe havens (and between them and developers of the traceability tools). Standards also exist for metadata (W3C DCAT) and for provenance information (W3C PROV), as we have discussed earlier, and should be adopted.

Beyond the identifiers, many other elements of such a pipeline are also already well established. In particular RO-Crate (RO-Crate, 2021), developed by the Research Object community, provides a common format based around the FAIR principles "to improve the potential for understanding and reuse of research by making sure that the information that is needed to make a published resource useful is associated with it, and shared as a whole" (ResearchObject.Org, 2021). A baseline for such a data pipeline would consist of a way to reference a bundle of research artefacts as a single entity and describe both what they are and their connection to one another (for example input data to be analysed, code for the analysis, and output data describing the results), and using metadata standards to uphold the FAIR principles while doing so.

A key goal of such a data pipeline is the provision of tools to allow researchers to manage data, outputs and their provenance, allowing them to keep track of files related to analyses they run, along with the specific versions of their code used to produce them, without extensive manual annotation.

An example, developed during the pandemic – the FAIR Data Pipeline (Scottish COVID-19 Response Consortium, 2021a) – provides a simple API to programmers in a variety of languages (currently R, Python, Java, C++ and Julia) that allows the pipeline to trace I/O by code using a simple wrapper around read and write calls, provided external input data is registered with the pipeline (via a separate API that provides metadata). It also allows deeper metadata provision for data in specified "internal pipeline" formats (e.g. tables, arrays) to allow better automatic

annotation and introspection of data, and an issue tracking system for attaching problems to data or code so that the provenance system can identify potentially 'at risk' research and policy outputs.

Metadata for data registered through these interfaces is available from a public, FAIR data registry (Scottish COVID-19 Response Consortium, 2021b), which also provides URIs to access all of the publicly available data. The FAIR Data Pipeline uses all of the standards referred to above to ensure that it does not reinvent the wheel, but also provides the benefits of local data management and remote metadata storage; immediate benefits that will hopefully encourage uptake. The pipeline operates fully offline when running code to allow it to operate in TREs or High Performance Computing (HPC) systems, though it ordinarily requires a remote connection to communicate with a remote registry and data stores when accessing data initially. Further work will be required to ensure that approaches like this can operate within a TRE or in conjunction with platforms like OpenSAFELY.

Recommendation: Develop pipeline(s) for FAIR data and metadata storage and management such as the FAIR Data Pipeline, and investigate mechanisms for integrating them into TREs and secure analytics platforms like OpenSAFELY.

3.3 Advanced data tools in data platforms

As illustrated in Figure 1, data platforms provide indispensable support to data-model-policy lifecycles (see Section 4). They are built by some data institutions and used by almost all (see Section 6). For the stakeholders of data-model-policy lifecycles, while technical skills are usually required to use data platforms, the provision of advanced techniques within data platforms can alleviate the shortage of particular skills as well as providing an effective means for skills training (Section 5).

The development of data science as an interdisciplinary field over the past decade has resulted in many different data analysis and data visualization techniques. There are many merits in having data platforms provide users with mature techniques through advanced data tools. As most data platforms are designed and developed by major data institutions, advanced data tools can be developed with a high-level of technical transparency and a rigorous process of quality assurance; can be cost-effectively shared by many users and enable knowledge sharing through data institutions; can facilitate the identification and promotion of best practices in working with data. Whilst it is necessary to develop the data skills of many users in order for them to use various data tools, poor availability of such tools, especially through trusted data platforms, is a common factor hindering the skills development of data users.

The experience of the current pandemic has indicated that many modellers have not had access to advanced data analysis and data visualization tools and do not have the time to search for, install, try, test and benefit from such tools. This topic is covered in more detail in (Chen et al., n.d.).

Recommendation: equip data platforms with advanced open-source tools for data analysis and data visualization. Such tools can significantly improve the effective use of data stored on the platforms and enable the integration of advanced data tools into data-model-policy lifecycles, whist providing hands-on content for skills training.

4 Data-model-policy lifecycles

Here we highlight the importance of the data-model-policy lifecycle that ideally links data collection with modelling, modelling with policy, and policy with intervention and back to active data collection. Here models represent any analysis that aims to extract useful understanding from data to inform policy options for effective intervention to control a pandemic including symptomatic testing, surveillance and contact tracing, which in turn can yield large amounts of data useful in modelling. We believe that the current paradigm emphasises a unidirectional flow from data to policy via modelling. This mitigates against feedback that would yield more efficient and effective gathering of the data needed to underpin interventions to control future pandemics.

We identify three challenges that represent significant opportunities to improve the data-model-policy lifecycle; (i) the level at which the data-driven analysis informing policy questions are communicated to decision makers, (ii) the infrastructure and tools for data collection and analysis along with relationships between data collectors and analysts, (iii) ensuring that insights and information from data-driven analysis and modelling influences policy to improve the active learning that modifies data collection during a pandemic. These challenges are likely to have parallels beyond public health policy.

4.1 Open Epidemiology standards

When using data to inform public health there is a pressing need to assess the quality of data and their suitability for the task at hand (Sections 1.1, 3). This includes the impact of biases, missing values and lack of precision (see (Swallow et al., 2021) Section 7.3). It is also critical to understand and track when these issues change for a given data source e.g., changes to the disease testing regime over time. Such concerns must also encompass the code used to process the data, and the models and analysis procedures used to analyse them as a data pipeline (Section 3.2).

A key challenge is therefore to develop approaches and accepted standards to assess the suitability of models and their outputs to inform policy questions including any likely ethical implications. It is critical to assess traceability of both model assumptions and data provenance and suitability. A standard but indispensable step is peer review of data and science quality including assessing suitability of data, the representation of epidemiological and other relevant processes, simulation and inference algorithms, and the limitations in each of these and consequent impacts on use. There is a welcome growth in the acceptance that software implementation of models and analysis needs to be both open and reproducible (O'Donnell, 2020). However, standards are needed to assess whether software works as intended with test results verifiable, repeatable and reproducible. Aspects such as automated code checking and implementation of regression tests should be routine and code should be written to be understandable, readable and documented for both users and developers (see for example a software checklist produced by Research Software Engineers (RSEs) for SCRC (Scottish COVID-19 Response Consortium, n.d.)). Model results need to quantify uncertainty related to imprecise knowledge of parameters and model structure. This should ideally be conducted through application of rigorous statistical tools for parameter estimation and model assessment ((Swallow et al., 2021) in this special issue).

The implementation of such standards will require improved data platforms, development of suitable skills and better data institutions and policies (6 and 7). However, even when carried out, the results of such assessments must be communicated effectively and in a manner under-

stood by end users i.e. the decision makers who will develop public health policy. This could, for example, be achieved by codifying the above assessments into an Open Epidemiology standard, communicated via scorecards that traffic-light and briefly summarise the results of each of the steps identified above. There are developments in this direction, for example through the RAMP initiative, the Scottish COVID-19 Response Consortium has developed the FAIR Data Pipeline (section 3.2) and a software checklist, which aims to incorporate all these aspects into a single open and traceable lifecycle.

Recommendation: Develop Open Epidemiology standards and scorecard systems that assess policy readiness – i.e. the suitability of models and their outputs to inform policy questions including ethical implications – underpinned by a) data and model traceability including quality and provenance b) assessment of data and science quality; c) open and reproducible software that is well documented and tested; and d) inference and model validation against simulated and real data including quantification of uncertainty.

4.2 Anticipating data and infrastructure needs

The variety, velocity and volume of data being collected during the SARS-CoV-2 pandemic has highlighted significant complexities associated with data collection. Novel viruses pose significant challenges in that there are often particular gaps in real-time understanding and current data available. There are challenges, both in terms of anticipating specific data that may be required (since the nature of the next pandemic will be far from certain), as well as anticipating the required infrastructures for collection, storage and dissemination of the associated data products. As already noted genetic sequencing, economic indices, mobility, contact and air pollution data have all been important information streams in the COVID-19 pandemic (Whitehead, Taylor-Robinson, & B., 2021; Travaglio et al., 2021). There is the potential for these to be routinely collected, even outside of pandemics.

Developing and maintaining open and effective communication between modellers and data holders through active collaboration will also assist with pandemic preparedness. Most data collection protocols have strict design protocols underpinning them for reasons of continuity, privacy and integrity. This means that generally the same data is collected in the same manner over time. There are benefits associated with this in ensuring continuity and the ability to detect trends without concerns of altered observation processes. New protocols often take considerable time to introduce, if they are possible at all. However, breadth and flexibility in data collection and analysis pipelines are vital to allow for variation in local circumstance and adaptation to changing needs. This is particularly important in low- and middle-income countries, where public health implications of epidemics are often severe, partly due to less developed infrastructure. During a pandemic data that can inform policy decisions as well as improve scientific knowledge is a priority. Improving the flexibility of data collection protocols in advance of future pandemics is therefore a significant but essential challenge that could improve and enhance model-data integration. A further challenge is to develop and implement guidelines for collection and curation of operational and happenstance data. With increasing development of new technologies, such as contact tracing apps, and sequencing procedures generating high-throughput genomic data, new data will become increasingly available to modellers and practitioners.

The second WHO report by the Independent Panel for Pandemic Preparedness and Response (IPPP, 2021) notes that the current global pandemic alert system is not fit for purpose. In many cases alerts are currently via news or social media. Platforms to collate epidemic intelligence from open and non-traditional sources have been created (WHO EIOS, 2021). What is needed

is an architecture that collects data in real-time from a distributed network of local clinics and laboratories with associated decision-making tools to provide early alerts and warnings. With smartphone availability reaching ever more of the global population there are significant opportunities in using the power of mobile apps to collect data relevant to pandemic and epidemic surveillance.

Linking these different sources of data and accounting for their dependencies is a significant challenge in terms of volume, privacy, granularity and their associated uncertainties. The ability to link between data sources depends significantly on having sufficient metadata available that allows models to reliably account for dependencies, as well as connections between patients/households/regions etc. to be made. There are additional privacy and anonymity constraints that are also at play here, which pose constraints on users of the data products in both access and publishing results, however there are ways of dealing with this (Tang, 2020).

The need to set up structures to facilitate communication between data collectors, modellers and policy makers is discussed further in section 7.

Recommendation: Communication between data collectors and modellers is paramount in supporting the production of robust data lifecycles and products and should be facilitated. Users of data should make the case for better linking between data products and data collectors should act on this by collecting and documenting appropriate meta-data on all products e.g. denominator population, observation processes, granularity, etc. Flexible and adaptive data collection procedures should be encouraged, as should the development of methodologies to inform these adaptive designs and exploit available data to the fullest extent possible.

4.3 Adaptive data collection during pandemics

The better linking of models and data provide a number of opportunities to improve use, collection and collation of data and especially operational data collected as part of ongoing activities or public health interventions. The analyses of model sensitivity and scenario projections provide ideal opportunities to study model deficiencies, as well as examine the information gain of varying data integration procedures (Pianosi et al., 2016; Jackson, Presanis, Conti, & Angelis, 2019; Simmonds, Jarvis, Henrys, Isaac, & O'Hara, 2020). Statistical parameter and model estimation, assimilation and inverse modelling allow existing data to improve model formulation, selection and integration (Swallow et al., 2021). In the process, these analyses can also help inform what data is required, through power analyses, uncertainty quantification and studies of parameter and model identifiability (Marion et al., 2021). The ability to refine and triangulate new sources of data informed by developing models in real time is a further closely linked challenge. There are also significant challenges for modellers and statisticians to develop appropriate methodology to make use of new and emerging sources of data that, while representing new opportunities, come with significant problems in terms of bias, coverage and interpretation, for example when modelling human behaviour (Galesic et al., 2021).

As knowledge of new viruses improves, and important transmission processes and comorbidities are discovered, new data may be needed, and tapping into existing infrastructure would be beneficial from both time and financial perspectives. To better exploit this modellers and analysts need to be more aware of the implementation constraints that could conflict with improved data collection. For example, priorities for diagnostic testing, primarily conducted to monitor rates of infection, do not currently maximise the value of such information for estimation of transmission rates and other epidemiological parameters (Swallow et al., 2021). However, the latter can be significantly enhanced, and conflicts between these goals minimised, by recording the characteristics of those being tested for both test-positive and test-negative

individuals. This reflects a general principle of data collection that information on who or what is assessed, in this case denominator population data, should be collected alongside primary data.

There are methodological challenges here in terms of estimation (Swallow et al., 2021), but also software and institutional challenges in developing suitable pipelines. Given the dynamic nature of policy needs during a pandemic (Marion et al., 2021) structured but adaptable feedback loops are needed between data collectors and modellers. There may well be useful lessons from the significant advances in information-driven data collection and automation for improving estimation in the environmental sciences (Fossum et al., 2019).

It is vital that decision makers are provided with clear and guidance as to what data would most improve situational awareness and understanding. Open Epidemiological Standards will help but communicating the benefits of enhanced data collection for effective decision making is paramount.

Recommendation: Value of information studies based on inference for epidemiological models should be conducted prior to and during pandemics and disease emergencies, and advice on the benefits of expanded and enhanced data collection should be communicated to decision makers in a way that makes clear the resulting benefits to control disease.

5 Data Skills

Modellers need to adopt an integrated, team-based approach to modelling in the future, defining the particular roles needed to allow computational work to deliver good data, insight and decision making in a pandemic (Czyzewski, 2021; Aguas et al., 2020). They need to move away from the "lone researcher" who-has-to-do-everything approach to a "team leader", who owns the scientific direction and selects the appropriate modelling approaches, but, depending on group size, skills and preferences, may write much of the code, write only prototype code, or simply work via discussion with RSEs and only contribute code within particular specialist areas.

The experience of the current pandemic has highlighted that, as we anticipate future pandemics, modellers also need to be more flexible in incorporating new datasets and running models on them – they need to adapt to reading and using different data sources. The adoption of more participatory approaches to data-model-policy decision making is evident in initiatives such as CoMo (Aguas et al., 2020) Critical to this is the recognition that skills in data curation and management are distinct from modelling and software development, and should be recognised in their own right.

While software engineering skills are beginning to be appreciated within the research community (although academic credit for such skills is still lacking), there is a dearth of data management expertise in academia, accompanied by a matching absence of respect in many disciplines for those critical roles that are tied to data generation and data management.

There are a wide range of data roles that could be envisaged as supporting this collaborative modelling enterprise:

- data managers, who set up and maintain data registries, and have overview of all data processes;
- data brokers, who can identify modellers' information needs, both proactively and responsively, and manage flow of data resources into and through the system;

- dataset experts, who understand relevant datasets, help identify useful data and the best choice of analysts, they might have a role in assigning quality metrics to data;
- data wranglers, who develop scripts to reformat and restructure datasets ready for use, and as part of the curation process;
- literature reviewers, who find and summarise relevant papers and reports;
- data analysts, who prepare data for analysis, produce data products and assign metadata, select and apply analytical methods, and create data visualizations;
- analysis experts who review analysis plans and data products;
- data mining experts, who recommend, develop, and test data analysis algorithms for supporting various analytical tasks, such as multivariate data analysis, time series analysis, classification and clustering, association analysis, network analysis, dimensionality reduction;
- visualization experts, who design visual representations for specific data, users, and tasks, develop visualization tools for monitoring and optimising the performance of individual models, and create visualizations for public dissemination (Chen et al., n.d.);
- data curators, who curate data resources sourced externally and convert them into required internal formats for the modellers;
- those who map data products to the structures required as input to a specific model.

Of course, some of these roles may be carried out by the same person, and some may be subsumed by the modeller, but all are important aspects of the process, and might usefully be specialised to some extent. There will be challenges in managing specialism, collaboration and recognition within traditional academic structures.

An important role, identified by the Scottish COVID Response Consortium, although only partially realised in practice, is that of 'data broker'. There are many datasets in existence, or which might potentially be collected, which might usefully inform modelling activities. This is particularly true of happenstance data (Section 3.1), which by definition may not have a wide prior audience, and/or may not even be collated until such time as someone identifies potential value. The existence of new data resources may even be sufficient to make viable modelling activities which would otherwise lack credibility. A successful data broker would have knowledge of available data resources, and a good understanding of their properties, their potentialities, and hence how they relate to the needs and objectives (both specific and more broadly) of the modelling team. Linking these together is not just an administrative task: done well, it will be a highly creative role. It is useful to explore the interaction between different roles. As listed above, the data broker will necessarily interact with the roles of data manager and curator, as well as with data analysts, visualisation scientists, and those engaging with the process of mapping data products to pipeline into a specific model. It is a key integrative role, whose opinions will be critical in prioritising choices when assigning staff resource to delivery of different data products. Note that the data broker does not necessarily have to carry out data analyses themselves, but probably does need to be able to see the potential for a particular dataset, when analysed in a particular way, to inform a specific requirement in a model.

The role of data analyst in a modelling project may be subtly different to that delivered by statisticians in other contexts. The focus has to be on delivering meaningful results, given the inadequacies and peculiarities of the available data, the skills readily available within the

group, and the time available. This is not an excuse for poor data management, bad choices in methodology or suboptimal analysis, but it does mean that there has to be a willingness to ensure that 'the best does not become the enemy of the good'. In this context, what becomes more vital is that the metadata for the data and analytic scripts are recorded appropriately, that the provenance of intermediate data products is clear, and that the metadata recording these issues is inherited by the model outputs. Good metadata is, however, also often the key to delivering better analyses in this context. Unless data has been collected specifically for a purpose at least analogous to that which it is being used in the modelling exercise, it is likely to exhibit bias. This is particularly true of happenstance data, but it also applies, for example, to the repurposing of administrative datasets. Metadata is vital in understanding what bias might be present, and the scope to use any associated covariates to statistically adjust for bias, using methods such as propensity scoring or multilevel regression with post-stratification.

Much could be learnt from looking at different collaborations during the current pandemic, what worked and didn't work in terms of data exploitation, to better understand what would help make collaboration easier. There are signs of increasing skills in the data science area and the nurturing the growth of Research Software Engineering and wider data disciplines. We need to understand what would help make this growth easier, drawing on case studies from the previous year to identify specific recommendations and ideas for wider mechanisms to improve the environment for doing these things well.

Recommendation: Data skills are critical to analysis, modelling and accurate policy advice, particularly during a pandemic, and Research Data Management and related skills should be promoted as an important aspect of research, just as Research Software Engineering has been over the last decade. Social science research should be commissioned to study research teams that were involved in the pandemic, both those that existed prior as well as those that were created de novo during the crisis. THese projects would review their working practices to establish how these groups worked, and failed to work, effectively, and how they managed different aspects of their work and how this relates to the wider data-model-policy lifecycle.

6 Data Institutions

Throughout the COVID-19 pandemic different modellers and their groups have been dependent on different data sources. We noted in section 1.1 that these different sources of data can lead to different predictions and model outcomes.

Across the world a wide variety of institutions have generated, maintained, made available and licenced the data that the modellers depend on. These organisations range from government departments to public service delivery bodies such as the various national health systems which operate independently but are politically accountable to their relevant governments. Data has also been provided by statutory bodies such as statistics authorities. These various organisations usually have the benefit of being publicly funded and are required, as part of their function, to capture and curate data.

Other organisations have generated, curated and made data available: universities, the private sector, NGOs, charities and Learned Societies. These have been important resources and one of the challenges we face is ensuring these various data resources are sustained, maintained and potentially consolidated going forward. Sustainability will ensure that institutions can steward data over the long term, in ways that help to increase the value that can be created from that data, while minimizing potential harms (Snaith, Szasz, Keller, & Tennison, 2020).

One concept that is attracting increasing attention is the Data Institution. These are organi-

sations that steward data on behalf of others, often towards public, educational or charitable aims. Data stewardship involves collecting, maintaining and sharing data, and, in particular, determining who has access to it, for what purpose and to whose benefit. Data trusts are one type of data institution. They provide independent, fiduciary stewardship of data (Open Data Institute, 2020). An example of an effective data institution is UK Biobank, established in 2004 as a registered charity, providing access to high quality biomedical health data, funded through a mix of grant funding and subscription.

Recommendation: The scope for an infectious-diseases-related Data Institution is one that should be seriously considered. It would serve as a first port of call for key data assets and would have a duty to curate and collate data from a broad range of sources. It could also be responsible for happenstance datasets such as those described in Section 1.3. It could be independent or be incorporated within an existing organisation.

7 Data Policies

Governments will simultaneously be the leading customer for insights generated from modelling activities, a significant generator of data of potential use by the modelling community (either directly by provision to access to government data or indirectly because of its ability to leverage cooperation from third parties), and the leading funder of most modelling work. It is important to assess whether there are aspects of government decision-making with potential to make these activities more or less effective. This topic is covered in more detail in (Hadley et al., 2021) but we will make a number of observations from the data perspective. Without appropriate data policies, FAIR data principles may not be implemented and other parts of the data ecosystem (Figure 1) may not function effectively or efficiently.

Particular policy challenges relating to data include:

- Incentives for data provision – are they (can they) be aligned to meet policy ends in a way that provides future flexibility? What should be the role of legal obligations and/or statutory powers for data provision?
- Legacy issues around data – do existing arrangements help or hinder data availability and access? What does better look like? How should we structure (and fund) ongoing capability?
- How can we make data accessible and useful under privacy constraints? What are the appropriate mechanisms by which you can modulate 'privacy' assumptions when you are in an emergency?
- How do we maintain a robust system for management of metadata in the context of mixed private and public data from multiple independent sources?
- How can we ensure that data collected for administrative or operational purposes is augmented by appropriate metadata that characterise the context in which individual data are recorded, e.g. who is tested and why?
- How can policies and support be put in place to ensure that happenstance data are used in a way that balances public benefit while managing concerns about individual privacy or business value?

The impact of all data and modelling oriented activities will depend on the ability of the customer to understand and make use of them. More fundamentally, the ability of modellers, statisticians and data scientists to orient their work towards policy-relevant activities, and to access vital data resources, will be constrained by the willingness and ability of government to understand the potential benefits from engaging early during a time of crisis, and to continue to engage between crises. In all such respects, the involvement of an informed customer will be vital. Alternative project delivery structures, where, for example, someone external takes on the role of project sponsor on behalf of government, thus helping maintain focus on genuinely impactful activities, are potentially valuable, but in turn these depend on the availability of a sufficiently informed third-party. The UK government and devolved administrations made good use of quantitative expertise, such as the Office for National Statistics, both to deliver key functionality and provide direction to external researchers, but it would be much preferable for relationships between policy staff and quantitative researchers to be managed and nurtured over the long-term, in advance of a crisis. It is to be hoped that initiatives such as the UK Joint Biosecurity Centre will be able to deliver aspects of this role.

Another useful exemplar for government interaction with modelling and data science comes from the structures developed to provide advice and expertise in the event of an incursion of an exotic livestock pathogen, such as Foot-and-Mouth Disease virus.

Since 2011 the EPIC (Epidemiology, Population health and Infectious disease Control) consortium has held a contract from the Scottish Government to provide on-going advice on the control and eradication of animal and zoonotic pathogens, and, in particular, to prepare to provide evidence to support government decision-making in the event of animal disease outbreak. The consortium comprises quantitative and qualitative researchers, epidemiologists and veterinary public health specialists.

Set up to implement aspects of the 2008 Scudamore Report on the 2007 Foot and Mouth Disease (FMD) outbreak, EPIC ensures that Government has access at all times to the expertise necessary to analyse information on animal movements and conduct risk analysis using the most up-to-date methodologies and techniques, EPIC scientists seek to develop and clarify the scientific advice and analysis government requires, both during times of disease freedom and during outbreaks in order to ensure appropriate information and analysis are available to enable evidence-based decisions to be made when assessing risk and implementing control strategies.

Substantial resource within EPIC has been devoted to the development of independent data management and curation functionality. Both formal and informal elements of data access management involve a substantial lead-time. A substantial time investment by data management and legal professionals is required when negotiating GDPR-compliant data sharing agreements (typically made more complex where data management has been outsourced by government, and/or where data use involves collaboration across multiple legal entities). Even the informal development of effective working arrangements to facilitate the movement and processing of data between data provider and data user, based on a shared understanding of the needs of the latter and the constraints on the former, will be a time-consuming process. It would be naïve to expect that strong and effective data management processes could be developed and implemented in a short period of time, and these would be even less likely to be achieved while negotiating the stress and conflicting demands of an outbreak emergency.

In addition, data collected for administrative purposes is subject to a rigorous curation process, ensuring that data is stored in well-defined, consistent and appropriate structures, having been processed using robust, tested data-processing scripts, and associated with appropriate meta-data that describes both the provenance of the data objects and additional quality-oriented meta-data where appropriate. Data providers have also had the opportunity to refine their

download procedures and scripts.

None of this infrastructure, whether hardware, software or intellectual capital, could easily or effectively be developed in real-time in the face of a disease incursion. The ready availability of pre-identified, pre-processed datasets, already familiar to key researchers, and very well understood by dedicated EPIC data managers, is now recognised by the consortium as a key asset. This level of preparedness is available only because of the willingness to fund the necessary activity prospectively over the medium to long term, not just reactively during crises.

A vital aspect of EPIC activities has been the role of policy/science brokers in supporting effective communication of the needs of government to quantitative modellers and in helping policy customers interpret quantitative and semi-quantitative results.

Another important leadership role within EPIC is best thought of as a type of 'project sponsor', where an experienced scientist with a sound understanding of the needs of government essentially maintains a watching brief to ensure and maintain the relevance of project outcomes to policy needs, helping fill any gaps in the capacity or capability of government policy staff.

This model of preparedness, with its emphasis on data provision and access, would serve as an excellent template for similar policy commitments in the field of human infectious disease control.

Recommendation: Develop policies and provide long term support to nurture relationships, expertise and improve dialogue and understanding between individuals and commercial, governmental and academic institutions to enable better interactions across the data-model-policy lifecycle. Preposition structures and resources akin to approaches like EPIC.

Discussion

If we are to tackle future pandemics more effectively with better and wider data assets at our disposal, we need to learn the lessons of the COVID-19 pandemic. We need to prepare and preposition skills and resources now. It requires an ability to collect, curate and the types of **data for modelling** described in section 1. It needs that data to be managed according to the **FAIR data** principles outlined in section 2. It needs investment in **data platforms** that support open data sets, access to sensitive linked data and trace provenance of model outputs as described in section 3. We need to recognize the complex **lifecycle of data** use and modelling, and support the technical and socio-technical processes described in section 4. This in turn will require the promotion and development of **data skills** and a recognition of the collaborative nature of effective data use in pandemic situations, described in section 5. It needs key **data institutions** to be sustainably funded at various regional levels, discussed in section 6. And **data policies** are required to enable full advantage to be taken of this prepositioned and continuously developing data ecosystem, as proposed in section 7.

References

Aguas, R., White, L., Hupert, N., Shretta, R., Pan-Ngum, W., Celhay, O., ... Coutinho, R. (2020). Modelling the covid-19 pandemic in context: an international participatory approach. *BMJ Global Health*, *5*(12). Retrieved from https://gh.bmj.com/content/5/12/e003126 doi: 10.1136/bmjgh-2020-003126

Apple. (n.d.). *Covid-19 – mobility trends reports – apple.* https://covid19.apple.com/mobility. ((Accessed on 07/27/2021))

Balabdaoui, F., & Mohr, D. (2020). Age-stratified discrete compartment model of the covid-19 epidemic with application to switzerland. *Nature Sci Rep*, *10*, 21306. Retrieved from https://doi.org/10.1038/s41598-020-77420-4

Bhaskaran, K., Rentsch, C., MacKenna, B., Schultze, A., Mehrkar, A., Bates, C., ... others (2020, December). Hiv infection and covid-19 death: A population-based cohort analysis of uk primary care data and linked national death registrations within the opensafely platform. *The Lancet HIV*. Retrieved from https://doi.org/10.1016/S2352-3018(20)30305-2

Blavatnik School of Government, Oxford. (n.d.). *Oxcgrt.* https://covidtracker.bsg.ox.ac.uk/. ((Accessed on 06/09/2021))

Bowman, V., Silk, D., Dalrymple, U., & Woods, D. (2020). Uncertainty quantification for epidemiological forecasts of covid-19 through combinations of model predictions. *arXiv preprint arXiv:2006.10714*.

Challen, R., et al. (2021). Risk of mortality in patients infected with sars-cov-2 variant of concern 202012/1: matched cohort study. *BMJ*, *372*. doi: https://doi.org/10.1136/bmj.n579

Chen, M., Abdul-Rahman, A., Archambault, D., Dykes, J., Slingsby, A., Ritsos, P. D., ... Xu, K. (n.d.).

Clark, J. (2021). Open science - a question of trust. *Data Intelligence*, *3*(1), 64-70. Retrieved from https://doi.org/10.1162/dint_a_00078

Coburn, B. J., Wagner, B. G., & Blower, S. (2009). Modeling influenza epidemics and pandemics: insights into the future of swine flu (h1n1). *BMC medicine*, *7*(1), 1-8.

CODATA. (n.d.). *Home - codata, the committee on data for science and technology.* https://codata.org/. ((Accessed on 06/10/2021))

CoMix study. (n.d.). *Comix study - social contact survey in the uk | cmmid repository.* https://cmmid.github.io/topics/covid19/comix-reports.html. ((Accessed on 06/09/2021))

Czyzewski, A. (2021). Modelling an unprecedented pandemic. *The Forum*. https://www.imperial.ac.uk/stories/coronavirus-modelling/. ((Accessed on 06/12/2021))

Data Gov UK. (n.d.). *Daily summary | coronavirus in the uk.* https://coronavirus.data.gov.uk/. ((Accessed on 06/09/2021))

DataCite. (n.d.-a). *Datacite metadata schema v4.4 mandatory properties.* https://support.datacite.org/docs/datacite-metadata-schema-v44-mandatory-properties. ((Accessed on 06/10/2021))

DataCite. (n.d.-b). *Datacite schema.* https://schema.datacite.org/. ((Accessed on 06/10/2021))

Davies, N. G., et al. (2021, April). 2021 estimated transmissibility and impact of sars-cov-2 lineage b.1.1.7 in england. *Science*, *372*(6538). doi: 10.1126/science.abg3055

de Haas, M., Faber, R., & Hamersma, M. (2020). How covid-19 and the dutch 'intelligent lockdown' change activities, work and travel behaviour: Evidence from longitudinal data in the netherlands. *Transportation Research Interdisciplinary Perspectives*, *6*, 100150. Retrieved from https://www.sciencedirect.com/science/article/pii/S2590198220300610 doi: https://doi.org/10.1016/j.trip.2020.100150

DOI Foundation. (2021). *Digital object identifier system.* https://www.doi.org/. ((Accessed on 06/10/2021))

Dong, E., Du, H., & Gardner, L. (2020). An interactive web-based dashboard to track covid-19 in real time. *The Lancet Infectious Diseases*, *20*, 533-534. Retrieved from https://doi.org/10.1016/S1473-3099(20)30120-1

eMBL-EBI. (n.d.). *The covid-19 infectious disease ontology < ontology lookup service < embl-ebi.* https://www.ebi.ac.uk/ols/ontologies/idocovid19. ((Accessed on 06/10/2021))

Epirecipes. (n.d.). *epirecipes: A cookbook of epidemiological models in r, python, julia (and more) - epirecipes.* http://epirecip.es/epicookbook/. ((Accessed on 06/09/2021))

Facebook. (n.d.). *Covid-19 interactive map & dashboard.* https://dataforgood.facebook.com/covid-survey/survey-and-map-data. ((Accessed on 07/27/2021))

Ferguson, N., et al. (2020). 2020 impact of non-pharmaceutical interventions (npis) to reduce covid-19 mortality and healthcare demand. *Imperial College COVID-19 Response Team*. doi: https://doi.org/10.25561/77482

Fossum, T. O., Fragoso, G. M., Davies, E. J., Ullgren, J. E., Mendes, R., Johnsen, G., ... Rajan, K. (2019). Toward adaptive robotic sampling of phytoplankton in the coastal ocean. *Science Robotics*, *4*(27).

G20 Leaders. (2016). *G20 leaders' communique hangzhou summit.* https://ec.europa.eu/commission/presscorner/detail/en/STATEMENT_16_2967. ((Accessed on 06/10/2021))

Galesic, M., de Bruin, W. B., Dalege, J., Feld, S. L., Kreuter, F., Olsson, H., ... van der Does, T. (2021). Human social sensing is an untapped resource for computational social science. *Nature*, 1–9.

GitHub GoogleCloudPlatform. (n.d.). *Github - googlecloudplatform/covid-19-open-data: Datasets of daily time-series data related to covid-19 for over 20,000 distinct locations around the world.* https://github.com/GoogleCloudPlatform/covid-19-open-data. ((Accessed on 06/09/2021))

Google. (n.d.). *Covid-19 community mobility reports.* https://www.google.com/covid19/mobility/. ((Accessed on 07/27/2021))

Google Dataset Search. (n.d.). *Dataset search.* https://datasetsearch.research.google.com/help. ((Accessed on 06/10/2021))

Hadley, L., et al. (2021). *Challenges on the interaction of models and policy for pandemic control (vsi future pandemics workshop).* Submitted to Epidemics.

Harris, J. K., Johnson, K. J., Carothers, B. J., Combs, T. B., Luke, D. A., & Wang, X. (2018). Use of reproducible research practices in public health: A survey of public health analysts. *PloS one*, *13*(9).

HDRUK. (2021). *Trusted research environments and data managemen: Past present and future.*

Heesterbeek, H., Anderson, R. M., Andreasen, V., et al. (2015). Modeling infectious disease dynamics in the complex landscape of global health. *Science*, *347*, 6227. doi: 10.1126/science.aaa4339

Hinch, R., Probert, W. J., Nurtay, A., Kendall, M., Wymant, C., Hall, M., ... others (2021). Openabm-covid19—an agent-based model for non-pharmaceutical interventions against covid-19 including contact tracing. *PLoS computational biology*, *17*(7), e1009146.

Imai, N., Dorigatti, I., Cori, A., Donnelly, C., Riley, S., & Ferguson, N. (2020). *Report 2: Estimating the potential total number of novel coronavirus cases in wuhan city china.* Retrieved from https://www.imperial.ac.uk/mrc-global-infectious-disease-analysis/covid-19/report-2-update-case-estimates-covid-19/

Imperial College London. (n.d.). *Real-time assessment of community transmission (react)*

study | faculty of medicine | imperial college london. https://www.imperial.ac.uk/medicine/research-and-impact/groups/react-study/. ((Accessed on 06/09/2021))

IPPP. (2021). *Second report on progress by the independent panel for pandemic preparedness and response for the who executive board, january 2021.* https://theindependentpanel.org/wp-content/uploads/2021/01/Independent-Panel_Second-Report-on-Progress_Final-15-Jan-2021.pdf. ((Accessed on 06/10/2021))

Jackson, C., Presanis, A., Conti, S., & Angelis, D. D. (2019). Value of information: Sensitivity analysis and research design in bayesian evidence synthesis. *Journal of the American Statistical Association*, *114*(528), 1436-1449. Retrieved from https://doi.org/10.1080/01621459.2018.1562932 (PMID: 32165869) doi: 10.1080/01621459.2018.1562932

Keeling, M., Moore, S., Dyson, L., Tildesley, M., & Hill, E. (2021). S1184_spi-m_university_of_warwick_road_map_scenarios_and_sensitivity.pdf. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/975911/S1184_SPI-M_University_of_Warwick_Road_Map_Scenarios_and_Sensitivity.pdf. ((Accessed on 06/10/2021))

Kerr, C., Stuart, R., Mistry, D., Abeysuriya, R., Rosenfeld, K., Hart, G., & et al. (2021). Covasim: An agent-based model of covid-19 dynamics and interventions. *PLoS Comput Biol*, *17*(7), e1009149. Retrieved from https://doi.org/10.1371/journal.pcbi.1009149

Klepac, P. a. (2020). *Contacts in context: large-scale setting-specific social mixing matrices from the bbc pandemic project.* Retrieved from https://doi.org/10.1101/2020.02.16.20023754

Marion, G., Hadley, L., Isham, V., Mollison, D., Panovska-Griffiths, J., Pellis, L., ... Villela, D. (2021). *Modelling: understanding pandemics and how to control them.* In preparation for Epidemics special issue.

Marrs, T. (2017). *Json at work: practical data integration for the web.* O'Reilly Media Inc.

Meadows, A., Haak, L. L., & Brown, J. (2019). Persistent identifiers: The building blocks of the research information infrastructure? *Insights*, *32*(1). doi: http://doi.org/10.1629/uksg.457

Mossong, J. a. (2008). *Social contacts and mixing patterns relevant to the spread of infectious diseases.* PLOS Medicine. Retrieved from https://doi.org/10.1371/journal.pmed.0050074

Navaratnam, A. V., Gray, W. K., Day, J., & Wendon, J. (2021, 4). Tim w r briggs, patient factors and temporal trends associated with covid-19 in-hospital mortality in england: an observational study using administrative data. *The Lancet Respiratory Medicine*, *9*, 397-406. Retrieved from https://doi.org/10.1016/S2213-2600(20)30579-8

Nextstrain. (n.d.). *Nextstrain.* https://nextstrain.org/sars-cov-2. ((Accessed on 06/09/2021))

Office for National Statistics. (n.d.-a). *Coronavirus (covid-19) infection survey, uk - office for national statistics.* https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/bulletins/coronaviruscovid19infectionsurveypilot/9april2021. ((Accessed on 06/09/2021))

Office for National Statistics. (n.d.-b). *Coronavirus (covid-19) - office for national statistics.* https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases. ((Accessed on 06/09/2021))

Office for National Statistics. (n.d.-c). *Deaths - office for national statistics.* https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/. ((Accessed on 06/09/2021))

Omori, R., Mizumoto, K., & Nishiura, H. (2020). Ascertainment rate of novel coronavirus

disease (covid-19) in japan. *International Journal of Infectious Diseases*, *96*, 673–675.

Open Data Institute. (2020). *Data trusts in 2020.* Retrieved from https://theodi.org/article/data-trusts-in-2020/

OpenSAFELY. (n.d.). *Opensafely: Home.* https://www.opensafely.org/. ((Accessed on 06/10/2021))

ORCID. (2021). *Orcid.* https://orcid.org/. ((Accessed on 06/10/2021))

Our World Data. (n.d.). *Coronavirus pandemic (covid-19) - statistics and research - our world in data.* https://ourworldindata.org/coronavirus. ((Accessed on 06/09/2021))

Overton, C. E., Stage, H. B., Ahmad, S., Curran-Sebastian, J., Dark, P., Das, R., ... Vekaria, B. (2020). Luke webb using statistics and mathematical modelling to understand infectious disease outbreaks: Covid-19 as an example. *Infectious Disease Modelling*, *5*, 409-441. Retrieved from https://doi.org/10.1016/j.idm.2020.06.008

Oxford COVID-19 Evidence Service. (n.d.). *Oxford covid-19 evidence service - the centre for evidence-based medicine.* https://www.cebm.net/oxford-covid-19-evidence-service/. ((Accessed on 06/09/2021))

O'Donnell, F. (2020). *Sharing models for covid-19: guidance and tools – the odi.* https://theodi.org/article/sharing-models-for-covid-19-guidance-and-tools/. ((Accessed on 06/10/2021))

Panovska-Griffiths, J., et al. (2020, 11). Determining the optimal strategy for reopening schools, the impact of test and trace interventions, and the risk of occurrence of a second covid-19 epidemic wave in the uk: a modelling study. *The Lancet Child & Adolescent Health*, *4*, 817-827. Retrieved from https://doi.org/10.1016/S2352-4642(20)30250-9

Pellis, L., et al. (2021). Challenges in control of covid-19: short doubling times and long delay to effect of interventions.
doi: https://doi.org/10.1101/2020.04.12.20059972

Pianosi, F., Beven, K., Freer, J., Hall, J. W., Rougier, J., Stephenson, D. B., & Wagener, T. (2016). Sensitivity analysis of environmental models: A systematic review with practical workflow. *Environmental Modelling & Software*, *79*, 214–232.

Public Health England. (n.d.). *Variants: distribution of cases data, 20 may 2021 - gov.uk.* https://www.gov.uk/government/publications/covid-19-variants-genomically-confirmed-case-numbers/variants-distribution-of-cases-data. ((Accessed on 06/09/2021))

Public Health Scotland. (n.d.). *Covid-19 daily dashboard | tableau public.* https://public.tableau.com/app/profile/phs.covid.19/viz/COVID-19DailyDashboard_15960160643010/Overview. ((Accessed on 06/09/2021))

Rattay, P., Michalski, N., Domanska, O. M., Kaltwasser, A., De Bock, F., Wieler, L. H., & Jordan, S. (2021). Differences in risk perception, knowledge and protective behaviour regarding covid-19 by education level among women and men in germany. results from the covid-19 snapshot monitoring (cosmo) study. *Plos one*, *16*(5), e0251694.

Read, J. M., Bridgen, J. R., Cummings, D. A., Ho, A., & Jewell, C. P. (2020). Novel coronavirus2019-ncov: early estimation of epidemiological parameters and epidemic predictions. *MedRxiv*. Retrieved from https://www.medrxiv.org/CONTENT/10.1101/2020.01.23.20018549V2

Rentsch, C. T., DeVito, N. J., MacKenna, B., Morton, C. E., Bhaskaran, K., Brown, J. P., et al. (2020). *Effect of pre-exposure use of hydroxychloroquine on covid-19 mortality: a population-based cohort study in patients with rheumatoid arthritis or systemic lupus erythematosus using the opensafely platform.* Lancet Rheumatology. doi: 10.1016/S2665-9913(20)30378-7

Research Data Alliance. (n.d.). *Rda | research data sharing without barriers.* https://rd-alliance.org/. ((Accessed on 06/10/2021))

ResearchObject.Org. (2021). *https://www.researchobject.org.* https://www.researchobject.org/. ((Accessed on 06/10/2021))

Riley, S., Walters, C. E., Wang, H., Eales, O., Ainslie, K. E. C., Atchison, C., ... Elliott, P. (2020). React-1 round 7 updated report: regional heterogeneity in changes in prevalence of sars-cov-2 infection during the second national covid-19 lockdown in england. *medRxiv*. Retrieved from https://www.medrxiv.org/content/early/2020/12/16/2020.12.15.20248244 doi: 10.1101/2020.12.15.20248244

RO-Crate. (2021). *Ro-crate metadata specification 1.1.1 | zenodo.* https://zenodo.org/record/4541002#.YMFUcDZKgkh. ((Accessed on 06/10/2021)) doi: 10.5281/zenodo.3406497

ROR. (2021). *Research organization registry.* https://ror.org/. ((Accessed on 06/10/2021))

Science Academies of the G7. (2021). *Data for international health emergencies: governance, operations and skills.* https://royalsociety.org/-/media/about-us/international/g-science-statements/G7-data-for-international-health-emergencies-31-03-2021.pdf. ((Accessed on 06/10/2021))

Scottish COVID-19 Response Consortium. (n.d., 2021). *modelling-software-checklist/software-checklist.md at main · scottishcovidresponse/modelling-software-checklist · github.* https://github.com/ScottishCovidResponse/modelling-software-checklist/blob/main/software-checklist.md. ((Accessed on 06/10/2021))

Scottish COVID-19 Response Consortium. (2021a). *Fair data pipeline.* https://fairdatapipeline.github.io/. ((Accessed on 06/10/2021))

Scottish COVID-19 Response Consortium. (2021b). *Scrc drams.* https://data.scrc.uk/. ((Accessed on 06/10/2021))

Simmonds, E. G., Jarvis, S. G., Henrys, P. A., Isaac, N. J., & O'Hara, R. B. (2020). Is more data always better? a simulation study of benefits and limitations of integrated distribution models. *Ecography*, *43*(10), 1413–1422.

SLS-DSU. (n.d.). *Introduction to the scottish longitudinal study.* Retrieved from http://sls.lscs.ac.uk/about

Smith, L. E., Potts, H. W. W., Amlôt, R., Fear, N. T., Michie, S., & Rubin, G. J. (2021). Adherence to the test, trace, and isolate system in the uk: results from 37 nationally representative surveys. *BMJ*, *372*. Retrieved from https://www.bmj.com/content/372/bmj.n608 doi: 10.1136/bmj.n608

Snaith, B., Szasz, D., Keller, J. R., & Tennison, J. (2020). *Designing sustainable data institutions [report] – the odi.* https://theodi.org/article/designing-sustainable-data-institutions-paper/. ((Accessed on 06/10/2021))

Song, W., Zang, P., Ding, Z., et al. (2020). Massive migration promotes the early spread of covid-19 in china: a study based on a scale-free network. *Infect Dis Poverty*, *9*, 109. Retrieved from https://doi.org/10.1186/s40249-020-00722-2

Sun, K., Chen, J., & Viboud, C. (2020). Early epidemiological analysis of the coronavirus disease 2019 outbreak based on crowdsourced data: a population-level observational study. *The Lancet Digital Health*, *2*, 4. Retrieved from https://docs.google.com/spreadsheets/d/1Gb5cyg0fjUtsqh3hl_L-C5A23zIOXmWH5veBklfSHzg/edit#gid=447265963

Swallow, B., Burgman, M., Challenor, P., Coffeng, L. E., Dawid, P., Goldstein, M., ... Vernon, I. (2021). Challenges in estimation, uncertainty quantification and elicitation for pandemic modelling.

Tang, D. (2020). *Decentralised privacy-preserving bayesian inference for mobile phone contact tracing.* Retrieved from https://arxiv.org/abs/2005.05086

The DELVE Initiative. (2020). *Data readiness: Lessons from an emergency.* Available from. Retrieved from https://rs-delve.github.io/reports/2020/11/24/data-readiness

-lessons-from-an-emergency.html
The Health Foundation. (n.d.). *Covid-19 policy tracker 2020 | the health foundation.* https://www.health.org.uk/news-and-comment/charts-and-infographics/covid-19-policy-tracker. ((Accessed on 06/09/2021))
The Lancet Respiratory Medicine. (2020). Covid-19 testing in the uk. *The Lancet Respiratory Medicine*, *8*(11), 1061. Retrieved from https://www.sciencedirect.com/science/article/pii/S2213260020304458 doi: https://doi.org/10.1016/S2213-2600(20)30445-8
The Royal Society. (2020). *Reproduction number (r) and growth rate (r) of the covid-19 epidemic in the uk.* https://royalsociety.org/-/media/policy/projects/set-c/set-covid-19-R-estimates.pdf. ((Accessed on 06/09/2021))
Thompson, R. N., et al. (2020). 2020 key questions for modelling covid-19 exit strategies. In *Proc. r soc. b. 287.* 287 20201405. (http://doi.org/10.1098/rspb.2020.1405)
Travaglio, M., Yu, Y., Popovic, R., Selley, L., Leal, N. S., & Martins, L. M. (2021). Links between air pollution and covid-19 in england. *Environmental Pollution*, *268*. doi: doi.org/10.1016/j.envpol.2020.115859
UniProt Org. (n.d.). *Uniprot.* https://www.uniprot.org/. ((Accessed on 06/12/2021))
University College London. (n.d.). *Virus watch | ucl institute of health informatics - ucl – university college london.* https://www.ucl.ac.uk/health-informatics/groups/public-health-data-science/research/virus-watch. ((Accessed on 06/09/2021))
W3C. (n.d.-a). *Data catalog vocabulary (dcat) namespace.* https://www.w3.org/ns/dcat. ((Accessed on 06/10/2021))
W3C. (n.d.-b). *The prov namespace.* https://www.w3.org/ns/prov. ((Accessed on 06/10/2021))
Wade, M., Lo Jacomo, A., Armenise, E., Brown, M., Bunce, J., Cameron, G., ... et al. (2021). Understanding and managing uncertainty and variability for wastewater monitoring beyond the pandemic: Lessons learned from the united kingdom national covid-19 surveillance programmes. *Earth and Space Science Open Archive*, 50. Retrieved from https://doi.org/10.1002/essoar.10507606.1 doi: 10.1002/essoar.10507606.1
Whitehead, M., Taylor-Robinson, D., & B., B. (2021). Poverty health and covid-19. *BMJ*, 372-376. doi: 10.1136/bmj.n376
Whittles, L., et al. (2021). Evaluating england's roadmap out of lockdown. *Imperial College COVID-19 Response Team*. Retrieved from https://www.gov.uk/government/publications/spi-m-o-summary-of-further-modelling-of-easing-restrictions-roadmap-step-2-31-march-2021
WHO EIOS. (2021). *Eios-coronavirus-newsmap.* https://portal.who.int/eios-coronavirus-newsmap/. ((Accessed on 06/10/2021))
Wikipedia. (n.d.). *Covid-19 pandemic on diamond princess - wikipedia.* https://en.wikipedia.org/wiki/COVID-19_pandemic_on_Diamond_Princess. ((Accessed on 06/09/2021))
Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., & Appleton, e. a., Gabrielle. (2016, March). The fair guiding principles for scientific data management and stewardship. *Scientific Data*, *3*(16001). doi: 10.1038/sdata.2016.18
Williamson, E. J., Walker, A. J., Bhaskaran, K., Bacon, S., Bates, C., Morton, C. E., ... others (2020). Factors associated with covid-19-related death using opensafely. *Nature*, *584*(7821), 430–436.
Williamson, E. J., Walker, A. J., Bhaskaran, K., et al. (2020). Factors associated with covid-19-related death using opensafely. *Nature*, *584*, 430-436. Retrieved from https://doi.org/10.1038/s41586-020-2521-4
Wong, A. Y., MacKenna, B., Morton, C. E., et al. (2021, January). Use of non-steroidal anti-inflammatory drugs and risk of death from covid-19: an opensafely cohort analysis

based on two cohorts. *Annals of the Rheumatic Diseases*, *19*. doi: 10.1136/annrheumdis-2020-219517

Wu., J. T., Leung, K., & Leung, G. M. (2020, February). Nowcasting and forecasting the potential domestic and international spread of the 2019-ncov outbreak originating in wuhan, china: a modelling study. *The Lancet*, *395*, P689-697. Retrieved from `https://doi.org/10.1016/S0140-6736(20)30260-9`

Wynants, L., Van Calster, B., Collins, G. S., Riley, R. D., Heinze, G., Schuit, E., & van Smeden, M. (2020). Prediction models for diagnosis and prognosis of covid-19: a systematic review and critical appraisal. *BMJ*, *369*.

YouGov. (n.d.). *Covid-19 public monitor | yougov.* `https://yougov.co.uk/covid-19`. ((Accessed on 06/09/2021))

Zhou, Y., et al. (2020). Effects of human mobility restrictions on the spread of covid-19 in shenzhen. *China: a modelling study using mobile phone data The Lancet Digital Health, ISSN:*, *2*, 417-424. doi: https://doi.org/10.1016/S2589-7500(20)30165-5