Ada Lovelace Institute

THE ROYAL SOCIETY

Expert review

Learning data lessons: data access and sharing during COVID-19

Findings from an expert workshop exploring lessons learned from data-driven initiatives that emerged in response to COVID-19.

Executive summary

The urgent need for evidence to understand and tackle the virus has made the pandemic a catalyst for data sharing, as public and private institutions came together around a common cause and worked at pace to contribute to the national need.

For example, the ONS Infections Survey – led jointly by the Office of National Statistics (ONS) and the Department for Health and Social Care with involvement from IQVIA and the National Biosample Centre in Milton Keynes, played a critical role in understanding COVID-19 infection spread across the UK. And the Rapid Assistance in Modelling the Pandemic (RAMP) initiative set up by the Royal Society attracted 18,000 volunteers and supported rapid-response, peer review of publications and new cross-sector modelling teams.

In summer 2020 the Ada Lovelace Institute and the Royal Society organised a roundtable event to discuss, under Chatham House rules, data-driven initiatives that emerged in response to COVID-19.

Participants with experience of a range of initiatives that had been rapidly setup in response to the pandemic, joined and shared lessons learnt from their experience and observation of these activities. They discussed the factors which enabled their success: a clear purpose and specific mission from the outset; confidence in public legitimacy; shared processes and data alliances; and clarity about the data landscape.

But the need for new collaboration also exposed the challenges around urgent data access. Common issues raised included: missing and poor-quality data; data access and need for data-sharing agreements; uncertainty about legal compliance and perceptions of legal barriers; and differences in methodology within multidisciplinary teams.

In the context of an emergency such as the COVID-19 pandemic, the speed in which the data is required makes issues of data discovery, cleaning, processing and lack of existing infrastructure particularly acute. But these learnings have application not only to times of crisis but to the broader National Data Strategy and the future of data governance in the UK and beyond.

In particular we see some fruitful lines of enquiry for better data sharing in the future that will repay consideration in 2021:

1. What standards, templates and legal guidance are needed to share data with confidence, for example:
    a. Does existing guidance sufficiently encourage private-sector and non-profit data-sharing when it is legal and in the public interest?
    b. Should Government provide data-sharing agreement templates to reduce the financial burden of legal advice and streamline data sharing, or does this pose risks or create unfair power dynamics?

2. How can cultural communication barriers be removed in multidisciplinary teams incorporating data scientists?
    a. What good models exist or conditions are necessary for effective working?

3. How should the public deliberate on innovation at pace?
    a. How can organisations feel confident they are working with public legitimacy?
    b. How do we understand what 'public benefit' looks like?

4. What is the right model for data matching and transparency about existing data?
    a. Do we need actors like data scouts to matchmake?
    b. What is the case for and against data registers?

5. Do we need to introduce a duty to share during emergencies?

The Ada Lovelace Institute and the Royal Society plan to take forward some of these questions in 2021 and would welcome registrations of interest via hello@adalovelaceinstitute.org.

Introduction

'Following the science' has been the rallying call of Government throughout its pandemic response. Using data has been a key mechanism by which 'the science' – the clinical, epidemiological, computational, behavioural evidence – has been followed and COVID-19 has demonstrated the need for swift and accurate data-sharing. However, enabling a range of actors to access, use and share data in response to the pandemic has involved a complex ecosystem of institutions, relationships, infrastructure, regulations and opportunities, with varying outcomes.

This data-driven crisis response has seen some significant public successes, in particular the study of factors associated with COVID-19 mortality on the openSAFELY analytics platform[1] and the European Commission-funded COVID-19 Data Portal.[2] And the Royal Society has launched three new initiatives using data for the COVID response: SET-C, RAMP and DELVE.[3]

However, the crisis response has also surfaced the difficulties of sharing relevant and timely data within and between sectors. Within the public sector, reports that local governments have been unable to access timely data on local outbreaks or shielded patients have raised questions about the technical, legal or political barriers preventing effective access to data.

Data sharing between the public and private sectors has also raised challenges. These range from direct questions of how much data Google and Apple's Exposure Notification System[4] should share with test-and-trace services about users of contact tracing apps, to social media providers' unwillingness to share commercially sensitive data for use in efforts to combat COVID-19 and 5G misinformation.

Issues around data sharing have had a long history and there has been no shortage of thinking in Government about what prevents data being used well. Since 1999, the topic of Government data sharing alone has, wholly or partly, been the subject of: three Government reports, an independent review, an 'open policymaking' process and a public consultation, two white papers, three acts, two codes of practice, and has provoked, in part at least, the creation of the Government Digital Service.

The government continues to explore new solutions, for example, giving the United Kingdom Statistics Authority the right to require the disclosure of information held by public-sector organisations, businesses and charities, for the purposes of carrying out its functions, in the Digital Economy Act 2017. The National Data Strategy, published in September 2020, is a culmination and continuation of this work.

---

1 Williamson, E. J. et al. (2020) 'Factors associated with COVID-19-related death using OpenSAFELY', *Nature*, 584(7821), pp. 430–436.

2 European Commission (2020) *COVID-19 Data Portal - accelerating scientific research through data*. Available at: https://www.covid19dataportal.org/ (Accessed: 8 December 2020).

3 The Royal Society (2020) *Rapid Assistance in Modelling the Pandemic: RAMP.* Available at: https://royalsociety.org/topics-policy/health-and-wellbeing/ramp/ (Accessed: 8 December 2020); *Royal Society convenes data analytics group to tackle COVID-19.* Available at: https://royalsociety.org/news/2020/04/royal-society-convenes-data-analytics-group-to-tackle-covid-19/ (Accessed: 8 December 2020); *SET-C (Science in Emergencies Tasking – COVID).* Available at: https://royalsociety.org/topics-policy/projects/set-c-science-in-emergencies-tasking-covid/ (Accessed: 8 December 2020).

4 Apple and Google (2020) *Privacy-Preserving Contact Tracing - Apple and Google.* Apple. Available at: https://www.apple.com/covid19/contacttracing (Accessed: 8 December 2020).

However, the COVID-19 pandemic has accelerated the need for new collaborations and in many cases shown there's still a long way to go.

To better understand these initiatives – their success and challenges – and to surface recommendations for better collaboration in the future, at the end of July 2020, the Ada Lovelace Institute and the Royal Society convened a roundtable with cross-sector stakeholders who have been involved in rapid data-sharing initiatives around COVID-19.

We then conducted five follow-up interviews with representatives from academia and government bodies involved in the initiatives discussed, to reflect further on multidisciplinary cultural barriers to data-sharing, differences across the nations of the UK, and collaboration within the public sector and between private and public bodies.

We are grateful to those who gave their time to participate in the workshop and those who agreed to help us explore some of these topics in greater depth.

New data access initiatives instituted during the COVID-19 crisis

Access to high-quality, relevant information plays an important role in identifying, understanding and solving problems. This is especially true when faced with a novel and rapidly spreading pathogen.

The pace of spread of the virus meant that Governments and societies needed to develop understanding of the threat quickly, and that decisions were taken based on data available in the moment.

Early understanding was based on first-hand accounts and anecdotal information decentralised across practitioners and patients around the world, rather than highly structured and well-understood existing datasets.

Countering the pandemic effectively demands that data is collected and shared widely and rapidly across institutions, sectors and borders, while ensuring data integrity and respecting the rights of those involved. Below, we outline eight examples of new UK data access and sharing initiatives discussed in the roundtable that were rapidly set up at scale in response to the pandemic:

DECOVID

A consortium of organisations came together to develop the DECOVID project, to bring together high-resolution, secondary care data from hospitals to track the unfolding of the COVID-19 crisis. Using frequently updated data, the project aims to provide researchers and clinicians with rapid insights leading to effective clinical, operational and regulatory strategies.

The project intends to minimise the burden of research data collection on frontline NHS staff, and collect near-real-time information on symptoms, test results, pre-existing medication regimes, patient movements and final outcomes. Statistical modelling and machine-learning techniques are used to analyse the data and inform clinical decision-making around best treatment options, timing of interventions and identification of risk factors.

The project has been developed through the collaboration of the Alan Turing Institute, University College London, the University of Birmingham, University Hospitals Birmingham NHS Foundation Trust and University College Hospitals NHS Foundation Trust. DECOVID will be placed within the infrastructure of PIONEER, the HDR-UK Health Data Research Hub for Acute Care.

ONS Infections Survey

The ONS Infections Survey is a major long-term study tracking the spread of COVID-19 in the general population.6 The survey aims to understand current infection and immunity levels (through antibody testing) to form an ongoing response to the outbreak.

The study forms part of Pillar 4 of the Government's COVID-19 testing strategy and uses a representative sample of the UK population.[5] The ONS expanded the survey to include Wales (in June), Northern Ireland (in July) and Scotland (in September), becoming the largest COVID-19 infection survey in the UK.

The ONS Infections Survey is led jointly by the Office of National Statistics and the Department for Health and Social Care, drawing on the expertise and research capabilities of the University of Oxford, data science company IQVIA and the National Biosample Centre in Milton Keynes.

5 Department of Health & Social Care (2020) *COVID-19 testing data: methodology note,* GOV.UK. Available at: https://www.gov.uk/government/publications/coronavirus-covid-19-testing-data-methodology/covid-19-testing-data-methodology-note (Accessed: 8 December 2020).

Project Odysseus

In 2017, the Alan Turing Institute initiated a project with the Greater London Authority and Transport for London to analyse air quality and design policy interventions using machine learning and statistical modelling.

In response to the COVID-19 crisis, the models, infrastructure and machine learning (ML) algorithms developed for the air quality project have been repurposed to monitor 'busyness' levels across London, with a team of researchers from the universities of Warwick, Cambridge and UCL.

Project Odysseus aims to analyse the public response to the interventions introduced to manage the crisis. The work is expected to generate insights to aid London's social and economic recovery, by identifying patterns of activity and behaviour across pre- and post-lockdown conditions.

The project integrates live data from multiple sources such as JamCam cameras, traffic intersection monitors and aggregate GPS activity from the air quality project, as well as point-of-sale counts and public transit activity metrics.

Vivacity and the Department for Transport

Data-driven transport optimisation company Vivacity has been supporting the Department of Transport to monitor urban mobility and examine the efficacy of social distancing.9

Vivacity uses an AI sensor network to capture and classify real-time mobility data from pedestrians, cyclists and motor vehicles. Performing various analyses on this data, Vivacity has identified levels of adherence to social distancing rules (finding for example that only 54% of pedestrians strictly follow 2m guidance). With these findings, Vivacity aims to help Government understand the impact of messaging and outcomes around social measures.

Rolls-Royce Emergent Alliance

Rolls-Royce has established the Emergent Alliance to facilitate data sharing between companies and organisations with the aim of responding to the challenges of COVID-19 recovery. The alliance is composed of over 60 companies supplying data and includes IBM, Google and Microsoft.

The Emergent Alliance has been working to combine a wide range of data such as business, travel and retail alongside behavioural insights to identify indicators of economic recovery and enhance decision-making for businesses and governments.

The Emergent Alliance highlights member initiatives such as Rolls-Royce leading the UK's ventilator challenge, which aims to increase manufacture and supply of ventilators to UK hospitals, as well as the Leeds Institute for Data Analytics' project enhancing modelling capacity for understanding lockdown exit strategies.

HM Revenue & Customs and the Driver & Vehicle Licensing Agency

HM Revenue & Customs (HMRC) and the Driver & Vehicle Licensing Agency (DVLA) have been working together on an identity verification project to support users applying for the Self-Employment Income Support Scheme (SEISS) using driving license data. SEISS requires customers with no prior online registration to sign up using the Government Gateway to claim their grants, a process involving an identity check that can be subject to delays and blockages.

To overcome such obstructions, for a possible 1.4m users, HMRC has extended the range of data that can be used for ID verification to spread the demand across more sources. In this effort DVLA has made license data available, increasing the capacity of the system and the chance of success for users.

RAMP

The Rapid Assistance in Modelling the Pandemic (RAMP) initiative was set up by the Royal Society to convene modelling expertise across disciplines to support COVID-19 modelling efforts.

The project has emerged from a motivation to support the academic community in pandemic modelling. Those with expertise were already overstretched between research commitments and policy advising responsibilities, through channels like Scientific Pandemic Influenza Modelling Group (SPI-M) reporting to the Scientific Advisory Group for Emergencies (SAGE).

In this effort, a call for volunteers was announced by the Royal Society, resulting in 1,800 responses. The project has quickly grown in scope, operating several smaller, targeted projects ranging from epidemic models to urban analytics and behavioural modelling.[6]

RAMP has continued its activities by implementing infrastructure and processes to enable crowd-sourced, rapid-response peer reviews of publications and datasets. With these activities, RAMP aims to inform the Government's work through supporting modelling efforts, establishing new research terms and triage and rapid review of literature.

PROJECT OASIS

Project Oasis is a third-party app management initiative, established between NHSX and the Ministry of Defence's jHub, to create coordination and coherence of the data collected through COVID-19 symptom tracking apps. The project aims to create a local and national-level understanding of how the virus is spreading.

This process involves the removal of any information that may identify users and ensures that only symptom and demographic data is transferred. The data is also checked for security issues and assured against NHS standards.

6 The Royal Society (2020) 'Urgent call for modellers to support epidemic modelling', 29 March. Available at: https://royalsociety.org/news/2020/03/Urgent-call-epidemic-modelling/ (Accessed: 8 December 2020).

Enabling factors for data-sharing initiatives

The discussions highlighted a number of factors which are necessary to the success of such initiatives. These include having a clear purpose, public engagement and public legitimacy, shared processes and data alliances and clarity on the data landscape and data custodians.

A clear purpose
The context of the pandemic brought a sense of shared purpose and mission between teams and partners across many of the projects. This was described as helping to smooth over misunderstandings, and promoting a willingness to trust that others were acting in good faith.

Project OASIS's success was partly attributed to having a small, focused team with a clear and specific mission from the outset; the shared purpose helped overcome the lack of pre-existing relationships.

Public engagement and public legitimacy
Leveraging existing citizen and patient engagement structures through NHS Trusts and charities was noted as a successful way of communicating the importance of data-sharing with patients and getting their input into shaping projects, particularly on privacy versus data-benefit trade-offs.

The DECOVID project, for example, through their Data Trust Committee and more widely, was able to use the Patient and Public Involvement part of Birmingham's PIONEER Digital Innovation Hub to understand patient and public thoughts on health data use and specific data access requests.

Participants also noted that having a clear signal from the public, either through active engagement or self-organisation in civil society, helps to smooth data-sharing, by reducing concerns about backlash in what are often risk-averse environments.

Project OASIS noted that two factors contributed to the success of the project: the desire for citizens to share their symptom data and NHSX being clear with citizens that they only wanted a specific, essential set of data.

Shared processes and data alliances
Data access was most successful when there were clear processes that provided information without the need to rely on personal contact, as well as offloading the burden from small private actors to the better resourced partner, particularly in the legal domain.

Project OASIS, for example, was conscious about getting data protection right up front, with responsibilities split between NHS England as the data controller and the Ministry of Defence as the data processor.

The project made it easy to onboard private symptom data providers, many of whom were small and spun up in response to the pandemic. To help the private providers avoid high legal fees for collaborating and give legal confidence to the smaller organisations, they provided a data processing agreement template for use with NHS England, written by lawyers but in a non-legal way.

Collaboration through alliances has also been a helpful way to speed up appropriate data access over the course of the pandemic in more sustainable ways. These alliances can allow for coordinated and confidential discussions to find out what is common to all the partners, and what is necessary and rational for data-sharing processes, while keeping secure and trustworthy data sharing.

Clarity on the data landscape and data custodians

Some respondents reported that Scotland health providers published a spreadsheet with metadata and descriptions of all the data available. This extensive catalogue meant those trying to access data knew what they would get and what process to follow, and so were able to access data more quickly than in other nations. Those reporting argued that this clear, well-explained process to apply and catalogue of assets made it more likely that researchers not already embedded in the health community would be able to contribute rapidly.

Data custodians have also been speeding up processes for COVID-19 data access to allow for timely release where possible, by redirecting resources towards more urgent requests and contracting staff to work longer hours. While this is not a sustainable long-term strategy, the ability to be flexible and prioritise resources towards the most urgent in an emergency is still an important enabling factor for successful data sharing.

The UK Health Data Research Alliance coordinated data custodians to streamline their processes and make datasets available through their Innovation Gateway. This has helped devolved nation's data custodians move towards a common data-access request form for COVID-related projects, which had been identified as an issue earlier in the pandemic.

Barriers to data access and sharing

While many of the initiatives laid out above have been successful in utilising data in supporting clinical practice, advising Government decision-making and relieving pressure on overburdened systems, those involved highlighted a wide array of legal, technical and cross-cultural challenges they faced along the way.

Difficulties in data sharing are not new problems and many of the barriers identified are not novel to the pandemic, although it has clearly exacerbated some existing issues and provided the impetus and urgency to overcome others.

The Centre for Data Ethics and Innovation's (CDEI) recent report *Addressing trust in public sector data* use provides a clear outline of existing challenges in the public sector. It examines where data has been shared between Government departments, and with commercial organisations, the barriers these efforts faced, and the steps that were and should be taken to address them.[7]

This report builds on the CDEI report and other existing work. Firstly, by giving a greater focus on cross-sectoral data-sharing between the public and private sector. And secondly, by examining these challenges in the context of an emergency, where the speed required makes issues of data discovery, cleaning, processing and lack of existing infrastructure particularly acute.

7 Centre for Data Ethics and Innovation (2020) *Addressing trust in public sector data use*, GOV.UK. Available at: https://www.gov.uk/government/publications/cdei-publishes-its-first-report-on-public-sector-data-sharing/addressing-trust-in-public-sector-data-use (Accessed: 28 October 2020).

Missing and poor-quality data

One of the main points emphasised through the roundtable was the pandemic's revelation about missing data. Death data and statistics have come under the spotlight in many of the debates and reporting following the outbreak, particularly in the context of the aggregated death statistics preventing an analysis of COVID-19's disproportionate effect on Black, Asian and minority ethnic people,[8] and through regulatory bodies withholding information on deaths occurring in specific settings such as care homes.[9] Through these cases, the crisis has provided demonstrations of the omissions in current statistical practices and data regimes.

Another respondent, focused only on publicly available data sources that did not require data-sharing agreements, had wanted to examine data by ethnic group across the country, but reported that both data availability and data quality across the nations made that challenging. Scotland's June report had significant gaps in the data, and confidence intervals were so wide that it wasn't possible to make inferences (this was acknowledged by the publishers of the data).[10]

Organisers of the RAMP project also reported difficulties accessing Intensive Care Unit bed data, which was available in Scotland but not in England. Several respondents echoed these comments in noting the differences between NHS England and Scotland, as well as the lack of standardisation across hospital trusts.

Issues of data quality and usability were discussed, particularly cases involving the collection and organisation of large amounts of data. The Good Sam app, a COVID-19 response initiative supporting the NHS delivered by the Royal Voluntary Service delegated tasks to volunteers, who were asked to deliver medicine to patients, take them to and from doctor's appointments and hospitals, and call them to prevent loneliness and isolation. While the NHS's call for help resulted in an impressive 750,000 volunteers, extensive work was required to extract valuable data and make this effort meaningful on a local level.

Data access and data-sharing agreements

One respondent explained that issues of data access are often not infrastructural, but dependent on knowing the right people. Getting access to certain datasets was only achievable through asking favours of researchers and relationships can affect the prioritisation of tasks. How to ensure the reliability of this process is an ongoing challenge, made more problematic by the time-sensitive nature of pandemic work.

---

8 Barr, C. et al. (2020) 'Ethnic minorities dying of Covid-19 at higher rate, analysis shows', *The Guardian*, 22 April. Available at: https://www.theguardian.com/world/2020/apr/22/racial-inequality-in-britain-found-a-risk-factor-for-covid-19 (Accessed: 28 October 2020).

9 Booth, R. (2020) 'Data on Covid care home deaths kept secret "to protect commercial interests"', *The Guardian,* 27 August. Available at: https://www.theguardian.com/world/2020/aug/27/data-covid-care-home-deaths-kept-secret-protect-commercial-interests (Accessed: 28 October 2020).

10 National Records of Scotland (2020) *Analysis of deaths involving coronavirus (COVID-19) in Scotland, by ethnic group*. Data up to 14 June 2020. Available at: https://www.nrscotland.gov.uk/files//statistics/covid19/ethnicity-deceased-covid-19-june20.pdf (Accessed: 8 December 2020).

The Data Evaluation and Learning for Viral Epidemics (DELVE) group described the 'landscape of data readiness', in which the governance mechanisms are in place to respond to areas of need at pace, alongside stewardship mechanisms to ensure that the response appropriately balances the benefits of data use with data subjects' rights. The DELVE work highlights the value of data from everyday transactions in understanding the pandemic and the impact of policy interventions. International examples show that, where past projects have established data sharing relationships, it enabled rapid but well-governed use of this data during the pandemic.[11]

Other respondents noted variable experiences in accessing and using data from across the different UK nations. One respondent noted that accessing English health data often relied on leveraging connections and extended conversations with the data controllers. Whereas Scotland had openly published lists of data that could be accessed along with clear instructions for how to apply for access on their website, reducing the need for connections and making the data more available to those newly examining health data due to the pandemic.

Data-sharing agreements were discussed as a process that can be difficult to navigate.

The effort of identifying vulnerable people, using free school meal eligibility as an indicator, made clear the unavailability of a single mechanism to facilitate data sharing across local authorities.

To be able to share this sensitive information with each other, 33 London boroughs would have had to establish 528 data-sharing relationships, and this barrier created 'unnecessary complexity, delay and security vulnerabilities.' The London Office of Technology and Innovation has recommended that boroughs make default use of the London Datastore and its team as a platform for data collaboration, in response to this problem.[12]

Another example given was the 'vulnerable people service', built by the Cabinet Office and GDS to identify people in need of specific services, such as prioritised supermarket deliveries. The GDS blog notes that the service 'enabled the delivery of over 4 million food boxes to clinically extremely vulnerable people by early August' and 'required collaboration between central government, local authorities and the private sector.'[13] This involved numerous data-sharing agreements between these parties. However, these have not yet been published. This lack of transparency poses a challenge to holding Government to account in its sharing of citizens' sensitive data with private providers.

11 The DELVE Initiative (2020), *Data Readiness: Lessons from an Emergency*. DELVE Report No. 7. Published 24 November 2020. Available at: https://rs-delve.github.io/reports/2020/11/24/data-readiness-lessons-from-an-emergency.html.

12 Hajri, G. (2020) Here's what we learned from two data collaboration initiatives during the Coronavirus crisis., *Medium.* Available at: https://medium.com/loti/heres-what-we-learned-from-two-data-collaboration-initiatives-during-the-coronavirus-crisis-98ae60a8e4ff (Accessed: 28 October 2020).

13 Ferguson, C., (2020) L*eading the digital, data and technology (DDaT) response to coronavirus*. Government Digital Service blog. Available at: https://gds.blog.gov.uk/2020/09/14/leading-the-digital-data-and-technology-ddat-response-to-coronavirus/ (Accessed: 28 October 2020).

Uncertainty about legal compliance and perceptions of legal barriers

Concerns about law and regulation were also repeated as a significant barrier to data sharing. Nervousness about compliance sometimes meant that data was not shared when it could have been, with one participant giving the example of a group of charities that opted not to publish data online due to apprehension about a past ICO enforcement.

While many organisations display great willingness to share data, a lack of experience in doing so, and legal uncertainty tends to be an inhibiting factor. It was indicated that template data-sharing agreements are helpful in overcoming these reservations.

Access to commercial data was identified as another challenge to be overcome, with one significant example coming from DCMS and their efforts to identify online disinformation being hindered by an inability to access commercially sensitive data.

DCMS set up a programme to counter disinformation with a dedicated analytics team looking across social media to flag concerning trends such as the burning of 5G masts. However, commercial blockages made platform access and data unavailable. DCMS has explored some of these issues in their Misinformation in the COVID-19 Infodemic Inquiry.[14]

Cross-disciplinary cultural differences

The cross-cultural divide between the hypothesis-free and expansive attitudes of data science and the minimisation principles in data protection legal practice was examined as a difficult tension.

One key difference was between traditional medical statistics and data science teams in their approach to analysis. The desire in data science to start off hypothesis free, visualising data, reducing dimensions and looking for signals conflicted with the health statistics approach, which outline exact tools and analysis beforehand.

This created difficulty in data science teams justifying data access needs to data controllers, who might be unfamiliar with their methodology and reasonably want to minimise data access. Data controllers typically wish to limit exposure to data and think in terms of access to personal data and minimising the number of people working on data at any one time.

Others highlighted that even when there were skilled data stewards able to manage data and provide rapid advice on information governance, there were often very few individuals within clinical data controllers capable of directing teams towards relevant data and guiding them through the access process.

Another difference highlighted was a lack of understanding of team size and composition needed for data science projects. Rather than a single 'data scientist' capable of handling an entire project alone, teams need both data engineers to build data pipelines and multiple analysts with bespoke skills, such as machine-learning trained statisticians not just statisticians with access to machine-learning tools.

14 House of Commons Digital, Culture, Media and Sport Committee (2020) *Misinformation in the COVID-19 Infodemic*. HC 234. Available at: https://committees.parliament.uk/publications/1954/documents/19089/default/ (Accessed: 28 October 2020).

Respondents made it clear that both traditional medical statistics and data science approaches were valuable in approaching problems. It was not a case of replacing existing talent in medical statistics within the healthcare system but developing mutual understanding with data scientist teams to complement their work, possibly through in-house data science expertise within clinical data controllers.

Conclusion

In this report, we have seen how much can be done at speed in crisis, the positives that we can take inspiration and learn from, and the many challenges that still prevent societies from using data to its fullest and in the public interest. Some of those challenges are exacerbated by the pandemic and other underlying challenges simply brought into sharper focus.

There are no easy answers to these challenges, but it is clear we need to have better models of data discovery and data sharing. Both to make us more resilient to the next crisis, and to take the lessons from this pandemic to improve data availability and access for tackling existing problems in human health and beyond.

Here we aim to set out some lines of enquiry we believe need further exploration to achieve that goal.

What standards, templates and legal guidance are needed?

- Should Government and large institutions provide data-sharing agreement templates to smaller partners, to reduce the financial burden of legal advice and streamline data-sharing?
  - Would this create an unfair power dynamic for partners unable to afford legal advice and so scrutinise those template agreements fully?
  - Can agreements be written in an accessible manner that allows low cost and trustable scrutiny by smaller partners?
- Does existing guidance sufficiently encourage private-sector and non-profit data sharing when it's legal and in the public interest?
  - Is there existing survey work examining what private-sector and non-profit organisations find most uncertain in existing guidance, and what they perceive to be the legal challenges to data sharing in the public interest?

How can cultural barriers be removed in multidisciplinary teams?

- What good models exist for cross-disciplinary teams incorporating data scientists?
  - Is there existing ethnographic research on cross-disciplinary teams to inform how teams are assembled in future?
  - Is there a need for specific ethnographic research looking at integrating data scientists into teams of statisticians?
  - Is this transferable across domains, e.g. from criminal justice to health?
- Is the NHSX Centre for Improving Data Collaboration best placed to consider how data science methods, approach and, teams can be better embedded in NHS trust settings?

How can the public deliberate on innovation at pace?

- How do we incorporate the public into purpose and agenda setting during a crisis, rather than just approving or disapproving already designed data access and sharing solutions?
- How do we understand what 'public benefit' looks like and bring citizens into those decisions?
  - In the health sector, do we need to distinguish between benefit for patients and benefits for citizens and wider society?
- Who should be responsible for ensuring that public engagement and deliberation happen during a crisis?
- What infrastructure needs to be in place to swiftly operationalise informed consultation?
  - Are there existing models or programmes that can be scaled up across academia and the health sector?
  - If not, what are good models from other sectors?

What is the right model for data matching and transparency about data?

- Do we need an explicit role and profession dedicated to matchmaking data need with data available? For example, 'data scouts' or 'data shepherds' who guide researchers and data users through systems to translate the data need into the data that is available.
  - Where should these roles be located?
    - » Inside universities?
    - » Inside data controllers?
    - » Inside organisations like the Health Data Authority?
  - Should this role also be explicitly tasked with meta-tagging data to reduce time and complexity for later data engineering?
- Should the focus instead be on better search engines and automated data discovery systems alongside more transparent, clear, open application processes that are self-explanatory?
- In what circumstances might one be favoured over the other? Or in what combination or order should they implemented?

- Do we need public-sector data registers?
  - What metadata should be standard across these registers?
  - What data should be available?
    - Should there be sample or representative synthetic data available to allow researchers to understand the data more fully before they request full access?
- Should there be a standard glossary of terms across public services to go alongside standardisation of data formats?
  - Would this make understanding the contents of datasets more accessible and increase comparability across data held by different providers?
  - Would this standardisation of terms instead obscure underlying differences in data collection and data quality, providing a false sense of equivalence that doesn't reflect the underlying reality?

Do we need a duty to share during emergencies?

- Should the Government create a duty to share for companies to allow access to the Office for National Statistics and Government departments during emergencies through the issue of compulsory data-sharing notices, modelled on the provisions in the Digital Economy Act?[15]
  - Is the definition of an emergency in the Civil Contingencies Act 2004 still a necessary and sufficient definition for this purpose?
- Who will provide independent oversight of data sharing that occurs, e.g., the ICO?
- Should there be mandatory reviews of the data sharing that occurred under these provisions once the immediate crisis has passed?
- Should we require the live transparency of data-access notices and data-sharing agreements made in these circumstances? Or in all circumstances?
  - Should this include, for example, publishing contracts and memoranda of understanding, in full and with any potential annexes, as soon as they are agreed and in machine readable format?

We propose to explore some of these questions ourselves in future work, but we would also like to see others explore these questions. If you are already working on these questions or have plans to do so, please don't hesitate to reach out to us at the Ada Lovelace Institute via hello@adalovelaceinstitute.org.

15 *Digital Economy Act 2017*, Ch 7. Statute Law Database. Available at: https://www.legislation.gov.uk/ukpga/2017/30/part/5/chapter/7 (Accessed: 28 October 2020).

About the Ada Lovelace Institute

The Ada Lovelace Institute was established by the Nuffield Foundation in early 2018, in collaboration with the Alan Turing Institute, the Royal Society, the British Academy, the Royal Statistical Society, the Wellcome Trust, Luminate, techUK and the Nuffield Council on Bioethics.

The mission of the Ada Lovelace Institute is to ensure that data and AI work for people and society. We believe that a world where data and AI work for people and society is a world in which the opportunities, benefits and privileges generated by data and AI are justly and equitably distributed and experienced.

We recognise the power asymmetries that exist in ethical and legal debates around the development of data-driven technologies, and will represent people in those conversations. We focus not on the types of technologies we want to build, but on the types of societies we want to build.

Through research, policy and practice, we aim to ensure that the transformative power of data and AI is used and harnessed in ways that maximise social wellbeing and put technology at the service of humanity.

We are funded by the Nuffield Foundation, an independent charitable trust with a mission to advance social well-being. The Foundation funds research that informs social policy, primarily in education, welfare and justice. It also provides opportunities for young people to develop skills and confidence in STEM and research. In addition to the Ada Lovelace Institute, the Foundation is also the founder and co-funder of the Nuffield Council on Bioethics and the Nuffield Family Justice Observatory.

About the Royal Society

The Royal Society is a self-governing Fellowship of many of the world's most distinguished scientists drawn from all areas of science, engineering, and medicine. The Society's fundamental purpose, as it has been since its foundation in 1660, is to recognise, promote, and support excellence in science and to encourage the development and use of science for the benefit of humanity. The Society's strategic priorities emphasise its commitment to the highest quality science, to curiosity-driven research, and to the development and use of science for the benefit of society. These priorities are:

- Promoting excellence in science
- Supporting international collaboration
- Demonstrating the importance of science to everyone

Find out more:

- royalsociety.org
- @royalsociety