The Alan Turing Institute

Workshop reports

Data science and AI in the age of COVID-19

Reflections on the response of the UK's data science and AI community to the COVID-19 pandemic

Below are the individual reports from the workshops that were convened by The Alan Turing Institute and the Centre for Facilitation in November and December 2020, following the Turing's 'AI and data science in the age of COVID-19' conference. The workshops were summarised by the facilitators and theme leads, and the editorial team of the primary report then applied light editing. These reports are a reflection of the views expressed by workshop participants, and do not necessarily reflect the views of The Alan Turing Institute.

Several of the reports end with references to extra case studies that were submitted by participants using an online form. Further details of these case studies can be found at the end of this document.

Part A: Public health, modelling and pharmaceutical interventions

Theme 1: Pathogenesis and virus evolution, vaccines and clinical trials (leads: Marion Mafham, Jim Weatherall)

Scope

The role of data science in bolstering the efforts of clinical scientists to analyse data at increasing scale and speed produced several potentially useful insights during the COVID-19 pandemic, from tracing the phylogeny of different strains of the virus to map its evolution, to predicting the structure of understudied SARS-CoV-2 viral proteins.

The UK response

The availability of a central data repository through the COG-UK[1] consortium generated whole genome sequence data (>100k sequences, November 2020), and was achieved through the coordination of numerous organisations (NHS, public health agencies, Wellcome Sanger[2] and academic institutions). Having those sequences available from the earliest stages, and the central coordination and prioritisation of platform trials, was crucial in pushing forward the development of vaccines, and will be vital for genomic surveillance as new strains of COVID-19 emerge.

The data science community repurposed and built on existing systems and databases (e.g. ISARIC,[3] ICNARC[4]). This adaptation lowered costs and enabled researchers-at-large to act more quickly. For example, ICNARC[5] processed and published extensive weekly summaries drawing on an existing audit that covered over 95% of intensive care units in the UK. Similarly, linked data from GPs and Trusts were available to researchers through the Discovery East London[6] platform that existed before COVID-19. Dashboards from this detailed local-level data fed into London and national daily situation reports.

Many members of the data science community collaborated and were ready to work outside areas of individual expertise to address the challenges of discovery and therapeutics, which helped speed research across institutes, disciplines, regions and demographics. One example of collaboration, combined with innovation and an open approach to data sharing, was the RAMP[7] initiative that allowed people to contribute to ongoing work and research. Another example of collaboration between organisations was the development of adaptive clinical trials (e.g. RECOVERY[8]).

Challenges

The ability of the data science community to respond more successfully to the challenge of the pandemic was hindered by several hurdles.

Better standardisation and documentation of clinical data (and other forms of data, e.g. location) would have been beneficial. Some positive examples of standardisation could have been replicated for far wider benefit. For example, workshop participants noted ISARIC as a leading example of effectively documenting existing data and providing clarity on the meaning of each data field, and the Phenomics[9] project as a good example of how standardised code lists could be expanded. Standardisation and codification of metadata would have made data more discoverable and interoperable.

Connected to the lack of sufficient standardisation, data pipelines could also usefully have been more readily accessible and linked. A good example of such a 'standardised' pipeline achieved through collaboration is OpenSAFELY,[10] which provides controlled access to a vast database of patient records to support COVID-19 data analytics across the UK healthcare population. However, this is only suitable for research studies where identifiable data are not required.

Further, workshop participants reported that greater access to healthcare data would have been beneficial. For example, much data sharing during the pandemic was undertaken using the COPI (Control of Patient Information)[11] notices as the legal basis. However, these notices were only ever envisioned to be a stop-gap measure. Consequently, consideration might be given to how the best aspects of COPI might be retained, whilst ensuring that the permissiveness does not undermine individual rights and protections. Otherwise, we may return to a context in which advances achieved during the pandemic would not be possible once the COPI regulations come to an end, despite evidence indicating public support for the use of health data in research.[12]

Workshop participants also reported that data sources could benefit from being better linked to each other – despite more collaboration, many UKRI-funded studies (ISARIC4C,[13] PHOSP-COVID[14]) are still relatively standalone. Connecting expertise – the right people to the right data to the right problem – was identified by workshop participants as a primary bottleneck in working on COVID-19's pathogenesis and virus evolution, and vaccines and clinical trials. This impacted the ability of the data science community to respond better, from sharing expertise and learning on a specific problem to avoid duplication, to including data asset holders, subject matter experts, researchers and clinicians in analysis to explain missing data or bias.

Shortcomings in the availability and detail of data inhibited the inclusion of patient subgroups in trials to ensure there was an appropriately diverse range of people involved. The lack of diversity amongst researchers carrying out work in this area was also raised by some in the workshop as a constraint on engaging vulnerable and underserved groups in research.

International context

The global nature of the pandemic led to numerous examples of collaboration and openness in which the UK's data science community contributed and benefitted, and which has led to further innovation in researching the pathogenesis and evolution of the virus. For example, the Protein Data Bank[15] enabled the sharing of data around protein structures and protein interaction. Moreover, COG-UK data and GISAID[16] data were also brought together in new genomic surveillance pipelines.[17]

1 COVID-19 Genomics UK (COG-UK) https://www.cogconsortium.uk
2 https://www.sanger.ac.uk
3 International Severe Acute Respiratory and emerging Infection Consortium (ISARIC) https://isaric.tghn.org
4 Intensive Care National Audit and Research Centre (ICNARC) https://www.icnarc.org
5 ICNARC COVID-19 report: https://www.icnarc.org/Our-Audit/Audits/Cmp/Reports
6 Discovery East London: https://www.eastlondonhcp.nhs.uk/ourplans/discovery-east-london-case-study.htm
7 Rapid Assistance in Modelling the Pandemic (RAMP): https://royalsociety.org/topics-policy/Health%20and%20well-being/ramp
8 https://www.recoverytrial.net
9 http://covid19-phenomics.org
10 https://opensafely.org
11 https://digital.nhs.uk/coronavirus/coronavirus-covid-19-response-information-governance-hub/control-of-patient-information-copi-notice
12 https://understandingpatientdata.org.uk
13 Coronavirus Clinical Characterisation Consortium (ISARIC4C): https://isaric4c.net
14 Post-hospitalisation COVID-19 study: https://www.phosp.org
15 http://www.wwpdb.org
16 https://www.gisaid.org
17 SARS-CoV-2 lineages platform: https://cov-lineages.org

Suggestions

- **Improve availability of, and linkages between, datasets**: It is vital to increase the availability of datasets in appropriate research environments to make it more straightforward to obtain and analyse in a much more connected way. The provenance of those data, their purpose, reliability, bias and what was left out, are also key to helping the data science community work effectively to better understand the COVID-19 pathogenesis and virus evolution and support the development of vaccines and clinical trials. However, given the limited usability of anonymous data, as part of this initiative the community must also address the issues of secure controlled access. All of these may be achieved by greater support for work being undertaken by Health Data Research UK (HDR UK)[18] and the UK Health Data Research Alliance,[19] who have an established role in connecting datasets and researchers and enabling healthcare data to be discoverable.
- **Increase preparedness through local resilience**: The experience of working with data at the local level across the UK was inconsistent, with some areas having far greater access to data and success working with it than others. At the local level, data needs to be available, and expertise and resources easier to find and deploy, so that local organisations have the autonomy to work with their data (as opposed to it being centrally held and decisions being made centrally).
- **Create an open data and clinical problem democratisation platform**: Following the example of Kaggle,[20] create a data-sharing and modelling platform that connects the right people to the right data for the right problem, ensuring it is integrated with links to existing data depositories while preserving privacy and security of the data.
- **Increase diversity and inclusion through better data analysis and communication of outcomes**: To monitor the efficacy of drugs and vaccines for subpopulations, the data science community needs access and governance frameworks to enable the appropriate and necessary collection, subgrouping and stratifying of relevant data. Better monitoring of demographics (age, sex, ethnicity of enrolled patients to vaccine and clinical trials compared to local eligible population) will highlight critical gaps.
- **Organise cross-agency 'war game' simulation** to connect expertise and identify critical data / ICT gaps in preparedness for similar pandemics.
- Recommend **broader and more rapid access to data**.
- **Coordinate with universities and industrial partners** to identify how to use data science to make a case report.
- **Encourage more cross-domain data linkage and modelling**, to explore what can be achieved at the intersection of these different datasets.

Useful resource

Software: Open Targets COVID-19 Target Prioritisation Tool.[21]

18 https://www.hdruk.ac.uk
19 https://ukhealthdata.org
20 https://www.kaggle.com
21 https://covid19.opentargets.org

Theme 2: Epidemiological modelling and prediction (leads: Spiros Denaxas, Deepak Parashar)

Scope

From describing transmission dynamics, to identifying risk factors for infection and adverse outcomes, and modelling control strategies, arguably epidemiology and data modelling have been at the core of the data science community's contribution to the COVID-19 pandemic response.

The UK response

The pandemic has accelerated academic research, with researchers focusing on providing rapid, accurate and actionable insights into the COVID-19 public health emergency. It has reshaped clinical academic research with rapid restricting and prioritisation taking place to accommodate the needed capacity on the NHS frontlines,[22] with different streams of work delivered in parallel and quick posting on public repositories of analysis protocol and code.[23]

Collaboration increased between different groups, as well as sharing of data and models. Examples include the collaboration between Bristol University and NHS planners to help project hospitalisation demand and bed occupancy,[24] and the approach used by SPI-M – drawing on the collective expertise of several infectious disease dynamics experts across multiple groups[25] – to deliver consensus guidance based on a comparison of outputs from multiple models.

Mathematical models of formal logic were used to capture prioritisation policy based on epidemiological and clinical choices with outcomes that are auditable and dependable. For example, the RAMP[26] and INI initiatives demonstrated the "step up" of the modelling community, with the Royal Society RAMP initiative modelling the entire potential outbreak (peak, second wave, etc.) at the start. Despite limited data, much early analysis has been reasonably accurate in predicting the burden of infection on clinically vulnerable populations, or estimating age-specific risks of COVID hospitalisation and death.[27] Novel approaches in tackling existing problems have been tried (e.g. triaging patients based on their condition rather than scoring has provided more granularity and discretion) and new ways of modelling and their application have been developed.

Data (other than for health settings) were collected from various sources, such as community surveillance schemes[28] and from the public providing information via apps (resulting in the COVID Symptom Study[29]). Data have also been collated from e.g. Facebook Data for Good[30], Google mobility reports which were used in calibrating models in terms of a proxy for social contacts,[31] and social mixing surveys.[32] All of these assets played an important role in forecasting the course of the pandemic.

Challenges

Problems in data access, availability, granularity and levels (micro vs. macro) posed important challenges to workshop participants. Often, work was initiated only for the team to be unable to progress due to a lack of data or a delay in accessing the necessary data, such as with research concerning the long COVID phenotype. Moreover, the ISARIC study[33] has been published but the underlying data have not been made widely available to researchers requesting access. In other cases, delays were encountered with accessing datasets when the purpose was data linkage across multiple sources in order to create longitudinal patient cohorts.

22 https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0237298
23 https://cmmid.github.io/topics/covid19
24 https://doi.org/10.1101/2020.06.10.20084715
25 https://www.gov.uk/government/groups/scientific-pandemic-influenza-subgroup-on-modelling
26 https://royalsociety.org/topics-policy/Health%20and%20wellbeing/ramp
27 https://eprint.ncl.ac.uk/270736
28 https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/bulletins/coronaviruscovid19infectionsurveypilot/4december2020 REACT: https://www.imperial.ac.uk/medicine/research-and-impact/groups/react-study/
29 https://covid.joinzoe.com
30 https://www.medrxiv.org/content/10.1101/2020.10.26.20219550v1
31 https://www.google.com/covid19/mobility
32 https://cmmid.github.io/topics/covid19/comix-reports.html
33 https://www.bmj.com/content/369/bmj.m1985

Another challenge was communication to the public and policy makers: the media misunderstood some model results (e.g. the Oxford study at the start of the pandemic) and some misleading statements received much media attention.[34] The advantages and limitations of modelling could have been disseminated to the public alongside the results, in order to enhance trust and enable understanding by a wider audience.

Suggestions

These fell into four areas:

Communication is key to ensure that the public, policy makers, health leaders and politicians understand and appreciate the data science community's information, benefits and recommendations. Workshop participants suggested that the Turing could better engage the community at different levels via its network of partner universities, and provide targeted "training" for leaders and policy makers.

The data science community should consider the information recipients need as "customers" when communicating with them, and use clear, transparent and reproducible communications. Notably, there are pre-existing mechanisms to support this type of communication, for example via the Science Media Centre,[35] which has made significant efforts to curate expert reactions to stories that are making the headlines in an effort to provide the necessary nuance and insight into the results.

Data are the cornerstone of effective epidemiological modelling. It is important to plan data sharing early as an essential part of any prospective data collection study. Effective future planning could be supported by:

- Having a directory of data sources together with guidelines for obtaining data access.
- Developing guidelines for establishing and enabling access to generic 'data lakes', containing anonymised data.
- Utilising expertise within the Turing to provide advice and guidance on the ethical and responsible use of data.
- Having a platform for data sharing and modelling for deeper modelling of COVID-19 data.

Inclusion and equality: Prioritise the exploration and understanding of the impact of COVID-19 on different ethnic and social groups and make sure that these insights are used in conjunction with other key risk factors (age, care home residency, healthcare professional worker status and other health exposures). This would help promote inclusion of underserved groups by researchers designing studies and by funders, reviewers evaluating focus areas, and teams delivering modelling projects. Further, increase the visibility of female and BAME scientists in the media.[36]

Methodology/collaboration: There were significant amounts of research waste identified in the COVID-19 literature, specifically where it concerned application of prediction modelling.[37] Clarify the capabilities of modelling and ensure access to the most reliable and accurate data by having a Turing panel for consultation on research, to save effort and time.

Workshop participants suggested that the Turing provides something similar to the Research Design Service,[38] which provides researchers with free advice on protocols for clinical trials (funded by NIHR), and that the Turing establishes a "Turing Health Research Unit" at its partner universities for a one-stop shop for advice on data access, methodology, and communicating to the public. This final point of engaging the public on research methods can increase public understanding and research credibility, thereby increasing public trust.

Extra case studies: 1, 5

34 https://www.medrxiv.org/content/10.1101/2020.03.24.20042291v1.full-text
35 https://www.sciencemediacentre.org
36 https://www.timeshighereducation.com/blog/women-science-are-battling-both-covid-19-and-patriarchy
37 https://www.bmj.com/content/369/bmj.m1328
38 https://www.nihr.ac.uk/explore-nihr/support/research-design-service.htm

Part B: Non-pharmaceutical interventions

Theme 3: Testing, contact tracing and other public safety interventions (lead: Mark Briers)

Scope

"Test and trace" is a well-established mantra of pandemic management that involves identifying, assessing, and managing people who have been exposed to a disease, and preventing onward transmission. However, understanding how to use modern technology and data science to facilitate this process has not been attempted at this scale before (i.e. prior to the COVID-19 pandemic).

The UK response

Multidisciplinary collaborations were quickly established to deliver impactful scientific contributions in a relatively short timescale. Exemplar publications were shared, helping to shape and inform policy.[39]

This collaborative process has improved modelling, data sharing, and interfacing to policy makers,[40] and has helped raise awareness about technology and data science for contact tracing. The data science community has developed methods to respond quickly to accelerate innovation, in an open and transparent manner, yet also allowing full formal peer review (for example, the DELVE working group).[41]

New data science related initiatives have been implemented, both as national programmes such as the COVID Symptom Study,[42] and localised programmes, for example the Southampton COVID-19 Testing Programme[43] and the Norwich Testing Initiative for universities.[44]

Challenges

Workshop participants highlighted that communication to the public could have been improved along several dimensions:

- Communicating statistical findings and data requirements: what data existed, the quality of the data, and the help that was needed.
- Communicating about how scientific data and models are informing policies, including success measures, evaluation frameworks, and the lessons learnt from pilots to support ongoing research.
- Communicating about how data science can promote positive health outcomes. This could have helped to achieve the objective of trust in data science (and models that underpin the analysis), which would have increased the acceptance of policies and led to a greater engagement. One example relates to public adherence of isolation notifications through test and trace; a targeting of the communication may have helped improve compliance.[45]
- Encouraging the public to engage with the contact tracing app.[46] Concerns about contact tracing focused heavily on technology and law and made it difficult to gain the public's commitment. If the public had engaged with the app, it may have helped to provide a clearer understanding about which venues and businesses needed to be closed to prevent the spread of the virus, rather than utilising national or city-wide lockdowns.

While data sharing has improved, more can be done towards data availability and visibility. Some hospitalisation data were not available in the early stages of the pandemic. The lack of a central repository for data made it challenging to establish what data exist, and this presented difficulties in rapidly connecting and using data.

Alongside data availability is the need to have clean data that preserve privacy and are ready to be analysed. Higher quality, spatial-temporal data could have assisted the support of policy at both a local and national level, specifically with the testing initiatives. Feedback loops around interventions could have been advanced with improved quality data.

39 https://www.medrxiv.org/content/10.1101/2020.07.13.20152439v1 ; https://www.medrxiv.org/content/10.1101/2020.05.14.20101808v2 ; https://www.ucl.ac.uk/health-informatics/groups/public-health-data-science/research/virus-watch
40 https://royalsociety.org/topics-policy/Health%20and%20wellbeing/ramp ; https://www.thelancet.com/journals/lanchi/article/PIIS2352-4642(20)30250-9/fulltext
41 https://royalsociety.org/news/2020/04/royal-society-convenes-data-analytics-group-to-tackle-covid-19
42 https://covid.joinzoe.com/
43 https://www.southampton.gov.uk/coronavirus-covid19/covid-testing/testing-programme
44 https://www.earlham.ac.uk/norwich-testing-initiative-covid19-testing-resources-universities
45 https://www.medrxiv.org/content/10.1101/2020.09.15.20191957v1
46 https://github.com/carstenmaple/SpeakForYourself

Participants suggested that the focus was too much on "today's" problem instead of proactively considering what the next likely problem on the horizon might be. Two areas were highlighted. First was the predictable risk of mass student migration to universities at the start of autumn term 2020, and the need to have a robust process for testing and contact tracing in place to deal with the consequent outbreaks. Second was the issue of how knowledge of immunity or vaccine status might be operationalised to support reduction in non-pharmaceutical interventions (e.g. the exploration of ethical and non-discriminatory immunity passports).

Workshop participants argued for greater use of better, effective and robustly designed apps and wearable technology for pre-symptomatic detection that can be used by the public, such as fitness devices and smartphones. For example, the existing app previously used for asthma patients to detect worsening asthma conditions could have been repurposed to detect COVID-19 by listening to the sound of coughing (even if app use may have been imperfect). Other technology improvements would also have supported the data science community, such as development of environmental (faecal/waste-based) tests across the sewage network, or air pollutant-based tests in public areas.

International context

Workshop participants believed there was notable divergence of approaches in the early stages of the pandemic, whereas a more effective response might have sought to collaborate with the international community who have had greater experience of dealing with epidemics.

Suggestions

Improved data sharing is necessary for a test and trace system to be effective. Collaboration across disciplines should be facilitated.

- **Establish 'data lakes'**: Workshop participants felt that cleaned, anonymised data ready for analysis would have been desirable, particularly in secondary care which currently is messy and difficult to access.
- **Maintain and further develop collaborations**: There are opportunities to further leverage the benefits that these collaborations have brought. Data scientists could take a more proactive role in communication to counter misinformation about testing, contact tracing and other public safety interventions.
- **Focus on forward thinking**: There needs to be more emphasis on foresight so that the next problem can be identified and addressed in a timely manner (e.g. considering vaccine administration while vaccines are being developed, and determining how data can be collected and made available as vaccines are rolled out to ensure the research opportunities are maximised). This also includes evaluation of previous policies, so we can improve testing, contact tracing and other public safety interventions.
- **Faster framework**: A robust research/analysis/review framework is needed that can be deployed in times of urgency so that delays in publishing ideas are reduced, and trust from the public can be increased. An example is an evaluation plan with objectives and assessment measures which can be used for public engagement with contact tracing.
- **Local approaches**: A localised approach to initiatives, such as contact tracing, may have increased levels of public trust and engagement.
- **Technology**: Workshop participants stated that testing and tracing technology needs to be improved, with greater development of detection at places where the public gather, rather than solely personalised detection. Two examples were highlighted: development of air quality testing systems at supermarkets and hospitals, and conversion of wearable based pre-symptomatic detection algorithms (e.g. through short ECG measurements with disposable pads or self-cleaning optical devices) into terminals placed at entrances to public areas, that are available for all to use. Notably, there was an appreciation that these tools have non-trivial ethical implications which would need to be addressed as well.

Useful resource
Leonelli, S. (2021). Data Science in Times of Pan(dem)ic . Harvard Data Science Review.[47]

Extra case studies: 2, 3, 5

47 https://hdsr.mitpress.mit.edu/pub/fi1rol2i/release/2

Theme 4: Behavioural analysis and policy interventions (leads: Tao Cheng, Ed Manley)

Scope

While 'Report 9'[48] may have brought the idea of policy interventions for managing a pandemic to the masses, the importance of non-pharmacological/behavioural interventions in this setting is well established. This highlights the role of data science in evaluating the effectiveness of policy interventions, such as lockdowns, distancing, masks, and fines for violating quarantine rules. It is vital that we understand what insights data science has contributed to the effects of interventions on behaviour, and how research on human behaviour has been captured in pandemic modelling, to identify what was done well, and to highlight opportunities for the near and distant future.

The UK response

A major breakthrough for the data science community's response was the opening up to data researchers of various mobility datasets by commercial providers (e.g. Cuebiq), alongside Apple and Facebook releases and the Google Mobility Report data. This enabled the detailed spatial analysis and large-scale simulation that was key to generating evidence for government and the public on the potential and actual impacts of lockdowns. Workshop participants cited public transport data from Transport for London and the Department for Transport as central to arguments about the effectiveness of the first lockdown. The ability to monitor data in real time through dashboards (e.g. i-sense COVID RED,[49] Evergreen[50] and the GOV.UK dashboard[51]) was also significant in the data science community's capacity to respond and report trends to the public at speed.

Behavioural modelling, especially in relation to mobility, was the standout area for success (in terms of practical application of behavioural analysis) in impacting policy intervention, enabling data scientists to advise and compare the potential consequences of different interventions (e.g. school closures, hospitality restrictions) for decision makers and the public to consider. Such analysis was key in emphasising the importance of lockdown periods to contain the spread of COVID-19 and prevent healthcare demand exceeding availability, while also underlining that other non-pharmaceutical interventions (testing and tracing, mask wearing, physical distancing, shielding, and self-isolation) on their own are insufficient.[52]

Workshop participants argued that there was a public appetite for clear explanations and visualisations of data. Some data journalists (in the UK and abroad) successfully communicated complex analysis and simulations to build public understanding of the pandemic, and the impact of specific interventions to 'flatten the curve'. Harry Stevens's article in The Washington Post featuring a 'corona simulator' is now the newspaper's most-read online story.[53] Moreover, John Burn-Murdoch's data visualisations helped explain 'excess fatality rates,' and illustrated how COVID-19 could not be reasonably likened to the flu. Further, regular and prolonged public engagement by prominent members of the data science community, such as Professor Devi Sridhar and Professor Christina Pagel, helped improve public understanding of the pandemic and policy interventions needed.

Challenges

The capacity of those working in behavioural analysis to contribute more to the response to the pandemic was impacted by numerous factors.

The most significant of these was a lack of transparency in terms of the use of data and studies in policy decision-making, data provision, and data collection methods. Many workshop participants observed that the lack of transparency in policy decision-making made it difficult to know which research studies had 'cut through' or even been considered by government and expert advisory groups when deciding on policy interventions. Policy makers were also not transparent about which data were important: it was known that Google and Apple data were central to policy-making, but the detail and quality of those data were unknown, while other data science (e.g. Faculty[54]) was not independently scrutinised.

48 https://www.imperial.ac.uk/mrc-global-infectious-disease-analysis/covid-19/report-9-impact-of-npis-on-covid-19
49 https://covid.i-sense.org.uk
50 https://www.evergreen-life.co.uk/covid-19-heat-map
51 https://coronavirus.data.gov.uk
52 https://doi.org/10.1016/S2468-2667(20)30133-X
53 https://www.washingtonpost.com/graphics/2020/world/corona-simulator
54 https://faculty.ai

There was a lack of transparency of apps used for monitoring – for example, the accuracy and effectiveness of the Test and Trace app remains unknown. This lack of transparency, combined with an array of different approaches by the various organisations involved in data collection and analysis, hindered the ability of the data science community to understand and integrate data from different sources.

Effective policy interventions often require localised responses (e.g. to implement lockdown at the fine spatial scale) but the data for localised analysis and modelling was simply not publicly available to do this. Supported by greater data sharing, cases for intervention could be strengthened by deeper modelling and combining local micro-level data with regional/national macro-level data. Currently, however, significant gaps in data availability means that modellers are forced to rely on assumptions, and the extent to which project findings stemmed from assumptions or data has not been made clear. The limitations of modelling were not widely understood, and it is vital to start communicating these as part of a wider drive to ensure confidence in modelling. Due to the lack of data for other non-pharmaceutical interventions (social distancing, mask wearing, self-isolation and shielding), the impacts of these interventions have not been fully integrated and calibrated in epidemiological modelling. There was also a lack of wider availability of certain data (e.g. financial transactions) at the granularity needed to understand behaviours and social change.

Despite some notable standouts, workshop participants felt that overall communication between the data science community and the public could have been better. There is an issue with poor communication of evidence, especially modelling outcomes, and a need to communicate uncertainty more clearly and effectively to a wider audience. Some workshop participants suggested that the lack of transparency had its roots in the reliance of decision makers on consultants who were able to provide data quickly at the start of the pandemic. Partly because of continuing to work with those consultants, decision makers' focus for behavioural analysis has remained stuck in 'fast-reaction' mode. As a result, we are failing to gain an understanding of the medium- and long-term consequences of non-pharmaceutical interventions and wider social behavioural changes.

The centrality of mobility, mobile device and other digital data to behavioural analysis has resulted in significant challenges around inclusion and diversity in the reporting and modelling of behaviour. Traditionally underrepresented groups are also scarce in digital data, especially those on the 'invisible' side of the digital divide and those that are not represented in any data.[55] Insufficient links between datasets meant underrepresented groups may remain underserved (e.g. the intersection of the spread of the pandemic with mental health or precarious labour markets). This is often made difficult where different geographical aggregations are used. Disadvantaged groups have been insufficiently engaged on what research would be useful for them, and the data science community would benefit from building on the work of the few organisations talking to these groups, such as the Ada Lovelace Institute.[56]

International context

The variety of global government interventions in response to the pandemic has provided a rich opportunity for data sharing and monitoring of the impact of those interventions, which can inform UK policy, and vice versa. A notable example of this is the Oxford COVID-19 Government Response Tracker, a comprehensive global tracking initiative from the University of Oxford's Blavatnik School of Government.[57]

There also appear to be areas where the UK could learn from the experience of behavioural analysis and policy intervention in other countries. For example, in Singapore and China, test and trace has been shown to work effectively only when linked with significantly more data, for example on mobility at the granular level of tracking the actual movements of those who are self-isolating. These approaches have rights implications (e.g. respect for privacy) that might not be acceptable in the UK, but merit detailed exploration to determine this formally.

55 https://data-feminism.mitpress.mit.edu/pub/h1w0nbqp/release/2
56 https://www.adalovelaceinstitute.org/our-work/themes/covid-19-technologies
57 https://covidtracker.bsg.ox.ac.uk

Suggestions

- **Develop more transparent models for predictions**: Provide mechanisms and techniques to give accessible and accurate interpretations to model predictions, so that both the computational process and the results of such process can be explained and trusted. Be more open about biases in data to avoid erroneous or biased data becoming the basis of intervention.[58]
- **Engage with training for public health professionals to better communicate data research findings**: There is a major need to communicate models effectively to decision makers and the wider public, and to link models to actions as a core part of the data science process; to debunk misuse of data; and to provide clarity on the aims of 'what-if?' modelling.
- **Improve access to a far broader spectrum of data**: For more effective analysis and modelling, including at the fine spatio-temporal level, it is vital to integrate other areas of data. The COPI notice approach that allowed short-term release of health data could be used for other data (e.g. sector data of relevance to understanding societal responses to policy).
- **Increase inclusion of underrepresented groups**: Research needs to investigate the medium- and long-term impacts of the virus. This will require more effective reaching out to communities but can also draw on the strong base of knowledge we already have on existing health inequalities and broader socio-economic inequalities.
- **Develop protocols and guidance on data quality** to ensure standardisation and transparency in data access, sharing and linkage.
- Lead public debate on the kind of **data related to behaviour that could and should be collected and monitored**, and the trade-off between privacy and public health, so that in the event of another pandemic the response can be mobilised rapidly, and the policy impact can be evaluated effectively.
- **Build strong connections with companies** to facilitate rapid data sharing in fine spatial and temporal granularity to enable local analysis for local policy-making.
- **Support provision and training in more data-intensive communication**: To gain the public trust needed for a major vaccination programme, it is vital to be able to present data in a way that is not prone to misinterpretation.
- **Build sustainable links between the data science community and others** – those who are closer to policy-making but also behavioural scientists who could enhance our interpretation of behavioural data.
- Enhance collaboration by emphasising an **interdisciplinary and data-driven research agenda**.
- Encourage research into **deeper and more integrated modelling** that goes beyond epidemic / transmission models.

Useful resources

Code: Stochastic age-structured model of SARS-CoV-2 transmission for UK scenario projections.[59]

Software: Identifying how the spread of misinformation reacts to the announcement of the UK national lockdown.[60]

Extra case studies: 2, 3

58 https://rss.org.uk/news-publication/news-publications/2020/general-news/professional-standards-to-be-set-for-data-science
59 https://github.com/cmmid/covid-uk
60 https://github.com/markagreen/misinformation_uk_lockdown_2020

Part C: Impacts

Theme 5: Non-COVID-related health impacts (lead: Bilal Mateen)

Scope

The COVID-19 pandemic raises health issues that go far beyond the immediate effects of the virus, and span an extensive range of physical and mental health conditions. As an example, during the pandemic, emergency admissions for cardiovascular disease plummeted, and several studies have been initiated to understand what happened to these individuals and whether (for example) we should expect a surge in admissions from late complications of missed cardiac events. Waiting lists for elective procedures are rapidly growing, screening services have been paused, and many oncological treatment pathways have been disrupted due to the fear of immunosuppressing patients whilst there is a highly virulent disease so prevalent in the community. Furthermore, reports have identified an increase in mental health issues and domestic violence.

The UK response

The 'COVID-19 priority' has sped the passage through ethical and data access procedures. This reduced bureaucracy has resulted in existing and new datasets being interrogated by emerging questions at a rapid pace[61]. It was recognised that the COPI notice has had a positive impact on the progress that has been made. Examples relevant to non-COVID health impacts include the BHF Flagship,[62] UK Biobank,[63] Health Data Research UK,[64] OpenSAFELY,[65] RTT performance[66] and characterising the shielded population.[67] The result of these efforts has been several studies that have already highlighted some non-COVID-related health issues that are a direct consequence of the COVID crisis. The London School of Hygiene & Tropical Medicine investigated the indirect acute effects of the pandemic on physical and mental health in the UK.[68] Notably, the largest reductions in GP contacts for acute physical and mental health outcomes during the pandemic were diabetic emergencies, depression and self-harm. Moreover, there are already early estimates of the huge backlog of elective surgery cancellations during the pandemic,[69] as well as on cancer services[70] and cardiovascular disease.[71]

Workshop participants commented on the fact that many analytics communities came together to share problems and best practice between healthcare systems to explore ways of meeting healthcare demands, including non-COVID health needs. Online forums such as FutureNHS[72] and various nationally run webinars have supported these practice sharing opportunities. More widely, new ways of working with data have been adopted, such as a two-way flow of data science research, e.g. working with police data to understand ethnic disparities and COVID-19.

Finally, a number of innovative solutions were quickly launched to address the loss of traditional means of population sampling that were no longer available during the pandemic. For example, the Mental Health of Children and Young People survey adapted existing methods to make more use of online methodology,[73] and these are now providing new longitudinal resources and supporting researchers to use different ways of population sampling beyond the pandemic.

Challenges

Whilst the UK was quick to respond to the immediate and direct COVID-19 health issues, as these could be well-defined, the full extent of the non-COVID effects remains uncertain. Some non-COVID-related health issues are already apparent. Wellbeing has been affected by lockdown measures.[74] Mental health deteriorated overall, but more for some groups,[75] especially the young, where one in five children with a possible mental health condition reported fear leaving home.[76] Participants agreed that for many non-COVID-related health issues it is still too soon to have a complete picture of the impact, but thought it would be a significantly greater impact than direct COVID. There is growing evidence that the medium-term impacts on diseases such as cardiovascular disease and cancer is extensive and that the longer-term impacts, including mental health, may be even greater still. As the boundaries of 'non-COVID health' are so unclear, this makes it challenging to curate the necessary data sources.

During the pandemic, many successes have been made through bilateral collaboration. Workshop participants saw the next step as more challenging, where collaboration involves more than two disciplines crossing disease boundaries. It was recognised that the lack of holistic understanding is more difficult and can result in outlining problems rather than outlining solutions and making decisions.

Whilst data access has improved as a positive consequence of the pandemic, workshop participants noted that there is more to be done. The challenges were associated with data that did not seem to exist, data that were not in the correct format, a lack of record-level linked data, being unable to access data, and data lags. A lack of standardised systems at healthcare institutions hinders the capture of decision-making-related data (e.g. patient triage). Having more live data dashboards for monitoring services, benchmarking, audit and feedback would have helped monitor the indirect impacts of COVID-19. Additionally, it was noted that a more streamlined and expedited ethics approval process would be helpful.

Data has different meanings, and the meaning of data has changed during the pandemic. As an example, if the number of heart attacks increases, is this due to patients' reticence to attend GP or hospital appointments face-to-face, or a lack of exercise during lockdown? Simple questions will often mean different things. An example is that the General Health Questionnaire, used by practitioners to support mental health screening, asks patients about their enjoyment in activities. The pandemic has changed the meaning of this question due to the restriction on many activities. The Referral to Treatment rules are another example of where current data may be irrelevant for our new times. Data are mainly collected for performance reporting against the 18-week target and now we are observing an unprecedented number of elective treatments, with waiting times in excess of 52 weeks.[77]

Tensions remain between quality and pragmatism, and between funding on urgent work with a short timeline to impact and longer-term studies. Convenience surveys with samples of unknown representativeness were noted to be potentially misleading,[78] especially because, unlike many routine/big data sources, there are often constraints on sample sizes which can restrict looking at relevant subgroups (e.g. ethnicity) using meaningful categories.

International context

We could learn from the Scandinavian model for data linkage that adopts an identity card and population register system to help data linkage, sampling, coherence, tracking and duplication. Normally, patient consent is not necessary, and research on health data does not need approval from a research ethics committee.[79]

Suggestions

- **Data sharing agreements**: Development of more general data sharing agreements as opposed to those just for very particular purposes/uses/users, to reduce barriers and delays. This is particularly important for non-COVID-related health issues as the impact is evolving slowly and the effects

61 https://www.imperial.ac.uk/media/imperial-college/medicine/sph/ide/gida-fellowships/Imperial-College-COVID19-NPI-modelling-16-03-2020.pdf
62 https://www.bhf.org.uk/for-professionals/information-for-researchers/national-flagship-projects
63 https://www.ukbiobank.ac.uk/learn-more-about-uk-biobank
64 https://www.hdruk.ac.uk/research/better-care
65 https://opensafely.org
66 https://www.tandfonline.com/doi/full/10.1080/17477778.2020.1764876
67 https://bmjopen.bmj.com/content/10/9/e041370
68 https://www.medrxiv.org/content/10.1101/2020.10.29.20222174v1
69 https://bjssjournals.onlinelibrary.wiley.com/doi/epdf/10.1002/bjs.11817 ; https://bjssjournals.onlinelibrary.wiley.com/doi/full/10.1002/bjs.11746
70 https://bmjopen.bmj.com/content/10/11/e043828.info
71 https://heart.bmj.com/content/heartjnl/early/2020/10/05/heartjnl-2020-317870.full.pdf ; https://www.medrxiv.org/content/10.1101/2020.06.10.20127175v1
72 https://future.nhs.uk
73 https://files.digital.nhs.uk/D1/D411D3/mhcyp_2020_meth.pdf
74 https://www.ons.gov.uk/peoplepopulationandcommunity/wellbeing/bulletins/coronavirusandlonelinessgreatbritain/3aprilto3may2020
75 https://www.understandingsociety.ac.uk
76 https://www.health.org.uk/news-and-comment/news/survey-presents-a-worrying-picture-of-children-and-young-peoples-mental-health
77 https://www.england.nhs.uk/statistics/statistical-work-areas/rtt-waiting-times
78 https://www.cambridge.org/core/journals/psychological-medicine/article/covid-conspiracies-misleading-evidence-can-be-more-damaging-than-no-evidence-at-all/4E2EB003C65CADB8C2C774B114D68BEF
79 https://ec.europa.eu/health/sites/health/files/ehealth/docs/laws_denmark_en.pdf

will be experienced in the longer term. The data science community could continue to promote this idea more, e.g. through public and policy maker engagement to help ensure long-term success of research in this area.

- **Record-level data**: A call to enhance the collection of routine, structured electronic health record (EHR) data. High-quality baseline data would have a more immediate impact. In a similar emergency pandemic scenario, it would provide a robust starting position to be able to evaluate the effect of the pandemic on health-related impacts and would have helped identify more easily non-COVID-related health impacts.
- **Address deficiencies in data sources**: This is the perfect opportunity to collect some routine, robust and structured non-COVID and non-communicable disease data. Participants agreed that capturing the missing or incomplete aspects of the NHS and allied services structured data is key to being able to understand non-COVID-related health impacts and inequalities, as missing groups are the worst affected. A widely predicted consequence of COVID-19 is a recession, and the physical and mental health issues are likely to be catastrophic, as is socio-economic deprivation. A proactive wide lens needs to be on 'missingness' such as missing people (homeless) and missing diagnosis (is there a fall in heart attacks or have they been undiagnosed?).
- **Holistic working**: Prioritise efforts that see data science as a route to bringing together multiple clinical disciplines (and economics experts). This may encompass better working with those on the frontline and planners who can highlight problems before they are noticeable in the secondary data.

Extra case studies: 1, 2

Theme 6: Economic and social impacts (lead: Karyn Morrissey)

Scope

The COVID-19 pandemic has touched every part of our lives. Beyond the devastating health effects, the pandemic and pandemic responses have caused the single largest contraction in the UK economy in the last 40 years. Moreover, COVID-19 has prompted procedural changes, such as replacing school-leaver exams with grade predictions. Many of these changes have disproportionately affected individuals from minority ethnicities and those from more socio-economically disadvantaged backgrounds. The data science and AI community may have played a role in mitigating or furthering these inequalities, for example with respect to the issue of data representativeness (for example, are minority groups appropriately represented in datasets used when modelling the effect of pandemic response measures?).

The UK response

Collection and timely release of data including mobility and socio-economic information was identified by the workshop participants as the biggest headline success. Mobility data released by private companies at the start of the pandemic was identified as particularly useful[80] and assisted with understanding how social contacts impact transmission. Local government and central government departments' data sharing supported young people and those who are vulnerable[81] and assisted in recognising the need for complete and accurate coding of ethnicity.[82]

Demand for disaggregated data by socio-demographic factors meant better breakdowns by ethnicity, income, sex, age, etc. to determine impact and identify where help is needed.[83] The contributions of the Office for National Statistics (ONS) were praised on a number of occasions in our workshops, including their work on deaths stratified by ethnicity,[84] as well as their analysis of the social impacts of COVID-19[85] and how people spent their time during lockdown.[86]

The necessary data on these key socio-demographic trends from the ONS and other public data providers were made available to the data science community, and frequently updated, which workshop participants believed made it possible for the community to inform policy makers and the public of the disproportionate impacts of COVID-19 on pre-existing disadvantaged groups. Furthermore, participants noted that the increased focus on generating social and socio-economic data was vital, rather than a sole focus on the headline economic impacts such as unemployment and GDP.

A further success highlighted was that members of the data science community were involved in the various cross-sector collaborations (academic, commercial, non-governmental organisations, etc.) that were rapidly formed in response to the pandemic. This included, for example, a unique collaboration between the data science community as part of the RAMP Urban Analytics group and Improbable (a UK technology company), to understand the impact of daily mobility and time use patterns and inform a COVID-19 hazard at the individual and small-area level. This made it possible to identify the differential impact of disease spread on vulnerable populations, such as those with severe mental illness,[87] and on ethnicity.[88]

Challenges

The first overarching challenge highlighted in the workshop encompassed difficulties that were encountered during the pandemic resulting from inaccessibility, availability and consistency of some data. Specifically, participants noted that there is a robust and rigorous system in place for obtaining epidemiological data (e.g. number of cases and deaths) via institutions such as Public Health England's daily updates. However, such a system does not exist for wider economic, mobility and socio-economic data, and this prevented workshop participants from understanding the wider socio-economic factors in the early weeks of the pandemic.

The visualisation of data via maps was a core means of communicating data on the pandemic to policy makers and the public. Data was generally made available at a local authority district level. However, to really understand the transmission of COVID-19 across communities, more granularity is needed, specifically with local-level data. This lack of spatial disaggregation, whilst preserving the privacy of economic and social impacts by demographic groups, was limiting and initially hid some social inequalities associated with the pandemic in its initial phase.[89]

The workshop participants noted that the data science community had better access to wider socio-economic data than industrial data, such as the impact of COVID-19 on production and employment by different industrial sectors. Participants questioned whether access to up-to-date industrial economic data at the local level would have supported the data science and AI community, particularly when using data to underpin lockdown scenarios. However, this data was not available, and lockdown decisions may have been made without being informed by data to properly understand the relevant economic and social context. This meant that the economic impact of proposed lockdown measures such as the publicly controversial 10pm curfew was not able to be measured via a robust data-driven cost-benefit analysis, weighing up both the health and economic impacts of such measures, as well as medium- to longer-term socio-economic impacts associated with such policies (such as loss of employment across the sector and subsequent impact on household welfare). This in turn meant that messages to both policy makers and the public were less clear and not as informative as they could have been, and the lack of transparency on the links between official sources of data and how policy decisions were taken resulted in public trust being eroded.

80 https://www.google.com/covid19/mobility
81 https://cdei.blog.gov.uk/2020/08/05/covid-19-repository-local-government-edition ; https://leedscc.maps.arcgis.com/apps/opsdashboard/index.html#/7a5fa9c4b3d74443be52d35dc1c34265
82 https://www.hdruk.ac.uk/news/championing-diversity-and-inclusion-through-data
83 https://fletcher.tufts.edu/news-events/news/disaggregated-data-and-real-reflection-covid-19-risks ; https://gisand-data.maps.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467b48e9ecf6 ; https://data.humdata.org/visualization/covid19-humanitarian-operations
84 https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/articles/coronavirus-covid19relateddeathsbyethnicgroupenglandandwales/2march2020to15may2020
85 https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/healthandwellbeing/bulletins/coronavirusandthesocialimpactsongreatbritain/4december2020
86 https://www.ons.gov.uk/economy/nationalaccounts/satelliteaccounts/bulletins/coronavirusandhowpeoplespent-theirtimeunderrestrictions/28marchto26april2020
87 https://journalofpsychiatryreform.com/2020/07/29/vulnerable-populations-differential-impact-of-covid-19-on-populations-with-severe-mental-illness
88 https://www.health.org.uk/publications/long-reads/how-to-interpret-research-on-ethnicity-and-covid-19-risk-and-outcomes-five; https://www.bmj.com/content/369/bmj.m2503
89 https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(20)32465-X/fulltext

A second challenge reported was the lack of a joined-up programme of data science activity at a national level across all sectors. This led to repetition within the data science community. Transdisciplinary collaboration was needed during this pandemic to bring together data and expertise from health, HM Treasury and the private commercial sector. Although there were some important developments in data sharing, there were instances where the culture of protecting data may have worked against the pandemic preparedness effort, both within the data science community and the private sector.

A third significant challenge that participants noted was that the data science and AI community lacked the necessary proactive approach to prevent and stop data misuse and misrepresentation. Effective communication with the public, media and policy makers could have mitigated against the many counter-messages and 'fake news' stories relating to the pandemic that emerged.

International context

The Oxford policy tracker[90] tracks and compares COVID-19 policy responses around the world, and the Imperial behaviour tracker[91] tracks global behaviours. Both proved to be useful resources when evaluating the UK response in an international context.

There may be something to learn from the trust gap in Black and Latinx communities in USA regarding the COVID-19 vaccine rollout.[92]

Suggestions

There was a strongly held view from the respondents to the survey and in the workshop that the COVID-19 crisis has exposed the vulnerabilities of individuals, societies and economies, calling for a rethink of how data on economic and social aspects of society are collected, shared and analysed in the data science and AI community. During the workshop, participants highlighted one all-encompassing suggestion in particular:

> *Focus attention on the medium- and long-term rather than just the immediate COVID-19 response, and build a data infrastructure that supports this.*

To support this suggestion, the following actions were proposed during the workshop. It was thought that these would help the community to better understand the socio-economic and wider societal impacts of COVID-19, including issues around inequality, by gaining access to appropriate health data and data on the complex determinants of health and using this in an appropriate manner. These actions would give the potential to develop large datasets that are safe and trusted by the public.

- **Standardisation**: Forming a public stewardship body which would develop and maintain a national data governance process with a standardised framework for data access to facilitate timely analysis. This body would be able to provide protocols and set a consistent approach for sharing private economic and social data prior to public health emergencies, with the benefit of addressing the need for more granular data identified in the challenges.

  The steps towards this ambition were identified as:

  1. The data science and AI community contributing to testing new data stewardship models.[93] The GO FAIR Principles[94] for scientific data management and stewardship would ideally be adopted.

  2. Encouraging companies that hold data to sign up for this.

  3. Holding conversations about privacy and putting agreements into place.

  4. Providing standardised indicators such as local economic performance and inequality, beyond the Index of Multiple Deprivation. Variables such as local economic value-added figures, gross local output, local unemployment rates and local housing affordability could be made public at an intermediary level of disclosure for the wider scientific community.

- **Collaboration**: The data science and AI community should make greater strides to work collaboratively and bring transdisciplinary groups together, including e.g. local communities, health and social care providers, local partners, analysts, third sector organisations, private organisations, and industry.

- **Public trust and privacy**: To enable the effective use of public data to address crisis situations, the data science and AI community must continue to build trust amongst the public. This can be achieved through a greater degree of transparency, as well as addressing public concerns related to data security, privacy and confidentiality, whilst simultaneously developing a better perception of the risk-benefit trade-off, for example in discussions of the benefits versus costs of lockdowns.

- **Public engagement**: Ideally, the data science community will develop vehicles to develop public trust by greater engagement with the public, for example by working with representative groups, informing them and consulting with them as the science around an incident such as the pandemic develops.

- **Rolling data collection**: Related to developing public trust and understanding in data science, rolling surveys to understand the economic and socio-economic impact of COVID-19 and various COVID-related policy measures (e.g. business closures) would ideally be undertaken to help fully understand the economic and social impact of such decisions.

Useful resources

Data and code: 'Living through Pandemics – Using Protective Cordons to Enhance Continuity.'[95]

Extra case studies: 4, 5

90 https://www.bsg.ox.ac.uk/research/research-projects/coronavirus-government-response-tracker
91 https://www.imperial.ac.uk/centre-for-health-policy/our-work/our-response-to-covid-19/covid-19-behaviour-tracker/
92 https://www.covidcollaborative.us/content/vaccine-treatments/coronavirus-vaccine-hesitancy-in-black-and-latinx-communities
93 https://theodi.org/article/applying-new-models-of-data-stewardship-to-health-and-care-data-report
94 https://www.go-fair.org/fair-principles
95 https://figshare.dmu.ac.uk/articles/dataset/Data_Code_for_Living_through_Pandemics_-_Using_Protective_Cordons_to_Enhance_Continuity_/12477515

Part D: Enabling data science response

Theme 7: Ethics, law and governance (lead: David Leslie)

Scope

The pandemic has brought into sharp focus several issues pertaining to the ethics and governance of data science, and AI research in general – from the impact of unrefereed scientific preprints and how this information is communicated to the public, to changes in research protocols in response to urgent demands for rapid response data analytics and AI innovation, to the permissive data sharing environment that was created by the COPI notice to facilitate the national response to the public health crisis.

The UK response

The view was expressed in our workshops that the experience of the pandemic might spur the development of broad-reaching guidelines for AI in healthcare. There has been considerable interest from many data scientists and some sections of the wider community in transparent and accountable data, scientific innovation, and Open Science. There has been particular interest in data sharing practices[96] and issues surrounding data accessibility and useability.[97] The government's Data Ethics Framework[98] and its public sector guidance for safe and ethical AI,[99] ONS's 'Five Safes' protocol for data integrity,[100] and the ICO and The Alan Turing Institute's guidance for explaining decisions made with AI[101] have all set standards for responsible data use and AI.[102] These cross-sector interventions in creating criteria for best practices for responsible data scientific research and innovation have helped to establish public expectations about "what good looks like" in socially impacting data use. They have provided a solid starting point for thinking about the development and codification of principles-based standards and binding regulations in situations of a global health pandemic, and in a post-COVID-19 world.

The widespread sharing of data was brought about by cross-sectoral and inter-organisational collaboration. There has been a willingness to collaborate in the context of the pandemic,[103] with new collaborative approaches, some of it stimulated by UKRI (and others) to generate more flexible and responsive modes of funding and cooperative, multi-institutional research, such as the DECOVID initiative.[104]

Some aspects of public engagement and participation have been positive;[105] work by the Ada Lovelace Institute[106] on how to engage the public rapidly and directly during lockdown has shown that, even amidst a public health crisis, the public's voice can be heard. Workshop participants felt that the data science community and wider groups of impacted individuals did "step up". For example, the Royal Statistical Society contributed to flagging up hazards early in the contentious Ofqual process.[107] Public dialogue surrounding the A Level grading issue had references to algorithmic bias and fairness and to data ethics and human-centred innovation, thus demonstrating that AI ethics has progressed some way to becoming a part of the moral vocabulary of public debate.[108]

Workshop participants highlighted the benefits of an Independent SAGE, which provided a model of how scientists can analyse, interpret and comment about the situation from a more evidence-based, neutral and interdisciplinary standpoint. The data community was also

96 https://opendatasaveslives.org
97 https://digital.nhs.uk/about-nhs-digital/our-work/nhs-digital-data-and-technology-standards/framework/beta---data-security-standards#the-data-security-standards
98 https://www.gov.uk/government/publications/data-ethics-framework
99 https://doi.org/10.5281/ZENODO.3240529
100 https://www2.uwe.ac.uk/faculties/bbs/Documents/1601.pdf
101 https://ico.org.uk/for-organisations/guide-to-data-protection/key-data-protection-themes/explaining-decisions-made-with-ai
102 https://www.gov.uk/government/publications/data-ethics-framework/data-ethics-framework-2020
103 DELVE: https://royalsociety.org/news/2020/04/royal-society-convenes-data-analytics-group-to-tackle-covid-19
104 https://www.decovid.org
105 https://covid.joinzoe.com/about
106 https://www.adalovelaceinstitute.org
107 https://rss.org.uk/news-publication/news-publications/2020/general-news/rss-alerts-ofqual-to-stats-issues-relating-to-2020
108 https://www.turing.ac.uk/blog/secret-life-algorithms-time-covid-19

responsive to the need from the public for regular project consultation and data updates, encouraging a more transparent approach to data sharing.

Challenges

The scramble to use existing datasets, or to rapidly create new ones recording COVID-19 symptoms and capturing clinical treatments and outcomes, ran the risk of proceeding without due regard for sampling biases. Workshop participants were concerned that biases may be "baked into" datasets (for reasons of systemic discrimination and structural inequalities), and other constraints intrinsic to data collection methods. Another concern was that AI and data science professional standards had been overlooked or neglected, and that existing recent guidelines and standards regarding AI were either not adhered to or not given full consideration, and that such derogation had been justified with reference to public health interests.

Urgency was often used as justification to avoid public engagement when setting up data collection and analysis. This sense of expediency and the race for outputs enabled a tendency to expand the claimed scope of the results of data analyses.

While the WHO has stated that the release of preprint results is a "moral obligation" in the context of public health crises,[109] the group agreed that the "post-truth" context of misinformation on social media and amplification of false news stories made best practices for publication during the pandemic complex. There is a lack of understanding of, and low/inappropriate involvement of, many data science communities in media processes.[110] Preprint protocols and accelerated peer review processes on papers must be approached with this in mind. And although paper retractions due to data inaccuracies, and commercial data sharing "by press release", seemed to be justified by "the pandemic", these corrective behaviours could also do damage due to the dynamics of social media message amplification, and the predominant culture of distrust surrounding data extraction and use.

A lack of pandemic-specific ethical oversight for many COVID-19 data projects is symptomatic of a wider lack of regulatory and statutory infrastructure that might effectively control irresponsible research and innovation behaviour, even as the changing environment of scientific challenges and practices provoked reflection.[111]

Over-reliance in the public health community on procurement and third-party vendors for data science and AI solutions was also highlighted. Many of the roles played by AI and data science in COVID-19 are filled by the increasing involvement of the commercial sector.[112] This creates issues around the tension between the 'public good' orientation of health-supportive AI technologies and the value profiles of private sector companies. It also creates issues around the systemic dependencies on 'Big Tech' for the provision of data processing infrastructure, off-the-shelf AI solutions, and other goods/services across digital supply chains.

International context

There was a strong suggestion from workshop participants that the focus of biomedical and public health related technologies should move from a national perspective to a more global viewpoint, with greater participation in transnational efforts of collaborative and equitable research and standard setting.[113]

International differences in classification and legal restrictions on collecting and sharing on matters such as ethnicity was highlighted, as it makes some data sharing difficult; for instance, US data includes 'Latinx', which is rarely used in Europe.

Suggestions

- Establish **professional accreditation**[114] for data scientists, including a code of conduct, with specific expectations for working rapidly/outside specialism in a humanitarian crisis, which would provide credibility and increase trustworthiness.
- Increase **public involvement** in evaluating

109 https://www.who.int/medicines/ebola-treatment/blueprint_phe_data-share-results/en
110 https://reutersinstitute.politics.ox.ac.uk/communications-coronavirus-crisis-lessons-second-wave ; https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2020-06/DNR_2020_FINAL.pdf
111 https://doi.org/10.1111/1740-9713.01463
112 https://reutersinstitute.politics.ox.ac.uk/industry-experts-or-industry-experts-academic-sourcing-news-coverage-ai
113 Such as the Research Data Alliance and the European Open Science Cloud
114 https://rss.org.uk/news-publication/news-publications/2020/general-news/professional-standards-to-be-set-for-data-science

outcomes, co-design in projects and negotiating overall priorities. This includes engaging with a wider audience, especially with minority and underrepresented groups, perhaps by setting up "citizen groups" or resourcing the Turing to be the focus of public engagement. Developing mechanisms for well-funded community involvement in setting research priorities and for the end-to-end participatory design of research and innovation projects was emphasised.

- Develop better ways to **communicate in an evidence-based and accessible way** and work with "today's world" of social media, preprints and the press, including increasing the understanding and involvement of data scientists in media processes. Communicate to the public better by providing channels for the provision of reliable information and inclusive means for critical information consumption.
- Form a **relevant body**[115] or use the Turing or a relevant national academy to bring together AI technologists, data science researchers, experts from across the disciplines of the social sciences and humanities, and members of civil society. The aim of this being to increase capacity across communities for understanding and operationalising responsible research and innovation practices in AI and data science in general, and with regard to the specificities that might arise in a global health pandemic. Utilise this body to provide the training and upskilling needed to produce ethical and democratically governed technologies.

Theme 8: Data readiness, collection and monitoring (leads: John Dennis, Sabina Leonelli)

Scope

High-quality data are the cornerstone of an effective pandemic response. The UK was able to rapidly create dedicated data assets to address issues related to the COVID-19 pandemic. The speed and complexity of the response, and deficiencies in existing infrastructures in some domains, meant that some aspects were neglected, and others not executed as precisely as necessary to be fully effective.

The UK response

Workshop participants believed that the AI and data science community had responded rapidly to the challenges of the pandemic, by supporting clinicians, local authorities and other organisations to collect and make use of all available data in innovative, iterative and collaborative ways. Workshop attendees commented about how data science work had come to "fruition" and that their contribution to the pandemic response had been welcomed and encouraged. The data legacy from the pandemic will be extensive, and offers the potential to support data scientists to respond rapidly in a future pandemic situation.

The pre-existing data infrastructure in the UK includes traditional data sources such as surveys and routinely collected patient data recorded in electronic health record systems. Open data provided by institutions such as the Office for National Statistics[116] and Public Health England have given researchers and the public the ability to access near real-time national data on COVID-19, including information on testing, deaths and hospitalisations.

There were regional data management systems that proved highly valuable. The SAIL[117] data linkage service was instrumental in redeploying existing datasets to support pandemic efforts in Wales. Public Health Scotland provided a daily data dashboard[118] to update policy makers and the public, including data collected from care homes. The 'DataLoch'[119] is a repository of all routine health and social care data for the Edinburgh and South East Scotland Region (funded by the Data-Driven Innovation Programme) and provides data to a range of researchers to address COVID-19-related questions.

As part of the data response, it was recognised that there were critical gaps in the data to support both research and policy-making. Funding was provided by UKRI, NIHR and many health charities to establish new datasets and management systems to address these shortfalls. Examples include the NHS COVID-19 Data Store,[120] the DECOVID dataset[121] (which aimed to provide a secure mechanism to collect data about patient care across all participating hospitals in the UK), and the Health Data Research Innovation Gateway,[122] established to provide information about the datasets held by members of the UK Health Data Research Alliance, facilitating the navigation of multiple datasets. The care homes Capacity Tracker,[123] established in 2019, was rapidly adapted to provide valuable information about infection monitoring and audit/compliance.

A major strength was that the UK already had in place a protocol for national collection of epidemiological data in the event of a severe respiratory pandemic. This was immediately deployed through the ISARIC4C[124] consortium, and meant researchers were rapidly able to provide vital information to characterise and prognosticate patients hospitalised with COVID-19, and to assess patterns of disease in children.[125] Respondents also identified the quality of data on critical care admissions provided in weekly reports by the Intensive Care National Audit & Research Centre (ICNARC[126]), and the utility of aggregate data on a similar critical care population collected by CHESS (COVID-19 Hospitalisation in England Surveillance System[127]) which was adapted from the UK Severe Influenza Surveillance System by Public Health England.

New collaborations were established to identify the prevalence of COVID-19 in populations. These include a partnership[128] between the University of Southampton, Southampton City Council and the NHS to develop a non-invasive saliva test for large-scale settings such as schools and workplaces. The COVID Symptom Study[129] supported the understanding of both disease spread and significant early symptoms. Another pressing need was care home data, and the VIVALDI[130] study sought to investigate levels of COVID-19 infections in homes.

Valuable support was provided by NHS England, the Open Data Institute[131] and several national disease societies who put in a huge amount of effort to facilitate access to existing data sources. These provided rapid answers to disease-specific questions around COVID-19 risk, including for rheumatic musculoskeletal diseases (see international context) and diabetes.[132]

Innovative solutions to provide real-time access, rapidly and securely, to detailed patient health records (notably the OpenSAFELY[133] platform led by the University of Oxford and London School of Hygiene & Tropical Medicine), were identified as likely to be of benefit to the UK health research base beyond the end of the pandemic.

Several UK longitudinal studies established specific data collections in response to COVID-19, including the UK Biobank[134] and Understanding Society.[135] Notably, the ELSA[136] study is collecting data on the effects of the COVID-19 crisis on the older UK population; data from the first wave of collection is available now from the UK Data Service.[137]

Local authorities and policy makers were able to make use of existing simulations produced by the AI and data science community, e.g. the SPENSER project (Synthetic Population Estimation and Scenario Projection Model[138]) and the COVIDSurg Collaborative,[139] which used global predictive data to inform surgical

115 UK Statistics Authority
116 https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/datalist?filter=datasets
117 https://saildatabank.com
118 https://public.tableau.com/profile/phs.covid.19#!/vizhome/COVID-19DailyDashboard_15960160643010/Overview
119 https://www.ed.ac.uk/usher/dataloch
120 https://www.england.nhs.uk/contact-us/privacy-notice/how-we-use-your-information/covid-19-response/nhs-covid-19-data-store
121 https://www.decovid.org
122 https://www.healthdatagateway.org
123 https://www.necsu.nhs.uk/capacity-tracker
124 https://isaric4c.net
125 BMJ 2020;370:m3339 / BMJ 2020;369 / BMJ 2020;370:m3249
126 https://www.icnarc.org
127 https://www.england.nhs.uk/coronavirus/wp-content/uploads/sites/52/2020/03/phe-letter-to-trusts-re-daily-covid-19-hospital-surveillance-11-march-2020.pdf
128 https://www.southampton.ac.uk/news/2020/09/saliva-phase-two.page
129 https://covid.joinzoe.com/data
130 https://www.ucl.ac.uk/health-informatics/research/vivaldi-study
131 https://theodi.org/topic/covid-19
132 https://www.thelancet.com/journals/landia/article/PIIS2213-8587(20)30272-2/fulltext ; https://www.thelancet.com/journals/landia/article/PIIS2213-8587(20)30271-0/fulltext
133 https://opensafely.org
134 https://www.ukbiobank.ac.uk/explore-your-participation/contribute-further/serology-study
135 https://www.understandingsociety.ac.uk/topic/covid-19
136 https://www.elsa-project.ac.uk/covid-19
137 https://www.ukdataservice.ac.uk
138 https://lida.leeds.ac.uk/research-projects/spenser-synthetic-population-estimation-and-scenario-projection-model
139 https://bjssjournals.onlinelibrary.wiley.com/doi/epdf/10.1002/bjs.11746

recovery plans. Local authority representatives from Richmond/Wandsworth and Haringey in the workshops reported how they now value predictive analytics, as these methods enabled them to identify residents likely to need additional support.[140]

Challenges

Workshop participants reported some difficulties which they believe impacted on effective data readiness, collection and monitoring.

The lack of a centralised data hub for all health data (not just NHS data) for the UK meant that it was difficult to avoid duplication of data. There were ethical/data protection limitations on access to data where they included personal data, and some respondents found it difficult to gain access to data because they were not health specialists.

Good-quality data requires systematic collection and curation, and this is extremely labour-intensive and expensive. Data collection methods in routine use within many healthcare organisations seemed neither appropriate nor robust enough for data collection on the scales required.[141] Respondents stated that data entry for most, if not all, COVID-19 epidemiological studies (e.g. ISARIC) were manual, and often could only be done by senior staff. Furthermore, each study had an entirely separate data specification and collection mechanism. The result of this was a substantial time drain for frontline hospital staff and a potentially detrimental impact on patient care.

The DECOVID[142] project had aspirational aims to address some of these issues and engaged the involvement of many data scientists, some of whom had limited experience of the health system. Workshop attendees reported that many found the realities of multi-dimensional healthcare data overwhelming, and this made it difficult to produce robust data analysis. An example is the simple question "Was the patient ventilated?". This may seem straightforward and intuitive, but given the complex treatment regime provided to many COVID-19 patients, in some cases it was impossible for clinicians to answer the question with a simple "yes" or "no".

Workshop participants reported that data collection for several crucial areas was either lacking, difficult to find, or else that there were inconsistencies in collection methods used across different regions which made direct comparisons of data challenging. Examples provided include:

- **Care home data**: Although data collection improved, initially there was limited information available and this meant some early signs of the egress of COVID-19 within the sector were missed.
- **Social care data**: There were no systems or incentives in place to support care agencies and personal carers to collect data about COVID-19 infections within home settings.[143]
- **School attendance data**: Respondents observed that they had established regular reports to inform policy makers using school attendance data. Some date input fields were removed with no warning or documentation to track the changes. This meant that the requested monitoring could not be completed.
- **Large-scale collection of detailed biological data** vital to understanding COVID-19 such as blood test results, proteomics, and genetic sequencing were reported by respondents to be lacking. Limitations of funding to projects like the BRAINS Imagebank[144] also meant that there were limited healthy controls to compare against brains that had possibly been damaged by COVID-19.

There were reported gaps in many data sources, particularly around standardisation of capture of socio-economic status and ethnicity, at both national and sub-national levels. Data standardisation would have enabled researchers to more easily triangulate information from different data sources and thus be better equipped to answer policy-relevant questions and to allow testing for any contextual effects on the outcomes.

Workshop participants also reported examples of where routinely collected data was perhaps not harnessed effectively by policy makers. An example of this was daily situation reports from individual hospitals, which provided information including daily bed occupancy rates for hospitals across England. This information could potentially have been used more effectively to manage patient intake and ease pressure on hospitals at or approaching unsafe occupancy levels.[145]

140 https://policyinpractice.co.uk/councils-get-faster-data-insights-to-boost-their-covid-19-recovery
141 https://www.gov.uk/government/news/phe-statement-on-delayed-reporting-of-covid-19-cases
142 https://www.decovid.org
143 https://www.health.org.uk/news-and-comment/blogs/strengthening-social-care-analytics-in-the-wake-of-covid-19-initial-findings
144 https://www.brainsimagebank.ac.uk
145 https://www.medrxiv.org/content/10.1101/2020.06.24.20139048v2

International context

The AI and data science community made use of global data sources and other data sources (e.g. the COVID-19 Data Repository[146]).

European cooperation was helpful, particularly where they were in a later stage of the pandemic to the UK. An example of a successful collaboration was between Newcastle University and the University Hospital in Modena, Italy. The hospital team provided UK-based researchers with data access and updates quickly, so that these could be used by the researchers to help with the predictions of likely respiratory failure in patients with COVID-19 pneumonia.[147]

The EULAR COVID-19 Database[148] was established to capture how rheumatology conditions and their treatment affected the risk of and severity of COVID-19. It was part of a global initiative that started from a post on Twitter. The team in the US shared a database template with the UK which enabled the EULAR team to go live less than one week later. Clinicians were able to quickly submit data and give feedback on the database design. Data import pipelines were set up with several EU countries. The UK database template has been shared with national societies in the EU and other disease-specific researchers to help them set up their own COVID-19 disease-specific databases.

Suggestions

- **Standardised data specifications**: There are clear benefits to creating stability in the data schema, and clarity in each field so that the perception of inconsistencies is reduced, and the data can offer improved real-time information with minimised subjective data processing. Changes in data collection and methodology should be clearly communicated, highlighting updates and changes in each data release. This will facilitate the comparisons between datasets and the sharing of data between different organisations, e.g. local authorities. Using a standard "baseline" data specification would permit linkage of various sources which would allow the testing for any contextual effects on outcomes. The standardised data specification should have a stress test by using regular expected events, e.g. the annual influenza season.
- **Automated data collection systems**: Data collection is resource-intensive, and there would therefore appear to be clear benefits to automated data collection systems, reducing the reliance on frontline staff to record data (often using manual systems).
- **Clear protocols for collecting data on protected characteristics**: These would include, for example, ethnicity, sex, age and socio-economic status. We should explore ways to develop clear protocols for anonymisation/synthetic data generation to include important demographic information in open datasets, such as the OpenPseudonymiser[149] technology developed by the University of Nottingham.
- **Investigate greater access to NHS data**: Access to high-quality, professionally managed, national NHS data repositories would clearly be of great value to many research groups. Clearly, this raises sensitive and complex issues around ethics and privacy, but given the very obvious research value of such data, there would seem to be some benefit to investigating ways to provide greater access.
- **Promote institutional and global data sharing for health emergencies**: Data sharing agreements between the UK and other countries take time to organise, and many were supported by EU funding. Having more flexible regulations for global data sharing during future health emergencies will help facilitate global learning.
- **Create a central repository of available shared datasets**, signposting to open datasets, and, for non-open data, the details of data sharing agreements and protocols for securely accessing datasets. The Turing could support the development of this type of 'data lake', enabling the provision of shared datasets and equitable data access, and facilitating unique linkages between datasets.

146 https://github.com/CSSEGISandData/COVID-19
147 https://www.medrxiv.org/content/10.1101/2020.05.30.20107888v2
148 https://www.eular.org/eular_covid19_database.cfm
149 https://www.openpseudonymiser.org

- **Initiate a strand of work on minimum data standards**, developing the existing international work on minimum standards and defining what metadata would look like, and explore the data requirements for decision makers in the UK.
- **Establish a forum to set principles and standards for data collection**, and promote the adoption of FAIR[150] (Findability, Accessibility, Interoperability and Reusability) principles, as well as responsible data collection and curation practices.
- **Transnational work**: Strengthen the global collaborations that evolved during the pandemic, e.g. the work of EULAR[151] and the Research Data Alliance,[152] to produce guidance for data documentation and metadata.
- **Develop collaborative working relationships between data scientists and clinicians and other domain experts**: Data scientists can provide insights about the collection and storage of data that clinicians may not be aware of. Clinicians and other domain experts can provide valuable insights to data scientists about the multi-dimensional nature of data collection within health and social care settings to improve collection methods and reliability across different settings.

Extra case studies: 1, 2, 3, 5

150 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4792175
151 https://www.eular.org/eular_covid19_database.cfm
152 https://www.rd-alliance.org

Extra case studies

Submitted by workshop participants using an online form.

Case study 1: Nottingham monitoring/ follow-up system

Organisation: Nottingham University Hospitals NHS Trust and the University of Nottingham.

Overview

We have created a near real-time system of monitoring and follow-up of all people suspected of having or confirmed as having SARS-CoV-2 attending Nottingham University Hospitals NHS Trust. We have developed and implemented an individual risk prediction score using all available electronic information (clinical observations, blood results, clinical decisions, oxygen and ventilation requirements and ongoing follow-up for readmission and vital status (in and out of hospital)). The prediction score is calculated on a daily basis using longitudinal information from the entire patient journey within the hospital. This system and score has been implemented in live hospital data and is being used to optimise resource and patient management within the hospital.

Who funded this initiative?

This was a collaboration of existing staff based in Nottingham University Hospitals NHS Trust and the University of Nottingham.

Overall, did the initiative make the desired impact?

Yes.

What worked well?

Excellent collaboration between all parties involved in the work. Rapid near real-time automatically generated, dynamic tracking, scoring and prediction implemented within hospital systems.

What were the difficulties and challenges?

Getting the time and resource to support it within the NHS.

Based on your learning from this initiative, what recommendations would you make for future best practice guides?

Continue to invest in using real-world NHS data on an individual level, and use the vast skills and knowledge within NHS Trust analyst communities to enable that to happen.

Case study 2: Discovery Data Service

Organisation: Barts Health NHS Trust

Overview

The Discovery Data Service (DDS) was developed in north east London with the aim of providing aggregated data from primary care and secondary care systems in a single system for point-of-care decision support, quality improvement initiatives, population health development and research. In the first instance, the data providers were GP practices in north east London and Barts Health NHS Trust.

Why funded this initiative?

At the beginning, the project was funded by the Endeavour Health Charity which is a charitable health trust endowed by Dr David Stables, one of the founders of the EMIS GP system. As the project has developed, contributions have been made through the NHS by STPs and the One London programme. Over the period of the project to date, NHS & charitable sources are matched.

Description / intended outcomes

The Discovery Data Service is a cloud-based publisher-subscriber system aiming to make access to large-scale health datasets from as many providers as possible. In this system, the data controllers remain the health organisations publishing the data to the DDS and data are only processed when appropriate data sharing agreements have been made between publishers and the subscribers to the system. Once in the DDS environment, the data are transformed to a common data model which in turn is defined by a semantic ontology. The data model is configured for business purposes, while the semantic ontology is configured for interpreting meaning using subsumption and reasoners.

The main intended outcome is to provide semantically useful and aggregated data both to publishers and subscribers of the data to support patients at the point of care, improve population health and to support researchers in developing new knowledge.

While the system is intended for large-scale data analytics, a specific use case has been developed to aid reporting of primary care assessments of COVID-19 infections during the pandemic. GPs record COVID-related structured data (SNOMED CT) in their clinical systems and this data can be linked to demographic information, hospital activity data (Admissions, Discharges and Transfers) as well as laboratory and imaging data. The DDS developed interactive dashboards in April 2020 which continue to be used at NHS London level to show COVID-related disease statistics in north east London, including geospatial displays of infection cohorts.

Overall, did the initiative make the desired impact?

Yes.

What worked well?

There are many examples of data extraction and analysis from the DDS. Three current examples illustrating point-of-care use, population health and research are:

1. NHS 111 frailty flagging for call handlers.
2. Primary care COVID tracking and reporting to NHS London based on primary care data during the pandemic (see above).
3. Provision of structured clinical phenotypic data for the East London Genes & Health programme, which aims to identify genomic health risk in the south Asian population of east London.

What were the difficulties and challenges?

1. Working with vendors to flow data into the DDS.
2. Data curation and data modelling for data taken from enterprise secondary care systems.
3. Securing data sharing agreements.
4. Securing agency support for the initiative.
5. Developing an agile but safe information governance process that includes patients and citizens.
6. Funding the DDS along with sustainable investment in health IT across sectors of health.

If the initiative did not achieve the intended outcomes, what do you believe were the reasons for this?

The DDS is in continuous development. Many intended outcomes have been achieved while others are still being worked on.

Based on your learning from this initiative, what recommendations would you make for future best practice guides?

1. Ensure each step in the development is defined by a project plan held in an overall programme.
2. Work early on the data access rules, data governance and citizen involvement systems.
3. Organise a programme board at the start of the project.
4. Explore different governance models for managing the project.
5. Invest and develop teams focused on data curation and data quality across systems as well as within publisher systems.

Case study 3: Nottingham overcrowding detection

Overview

Overcrowding detection for city centres and shopping malls.

Who funded this initiative?

Supported by Nottingham City Council.

Description / intended outcomes

The aim was to collect footfall and people count data (using our people count devices) to detect overcrowding in Nottingham City Centre using spatial analysis and clustering techniques. Based on the output algorithms, an alert is sent to venue managers.

Overall, did the initiative make the desired impact?

No.

What worked well?

The development of data collection tools and data analysis algorithms and the novelty of the approach.

What were the difficulties and challenges?

The City Council didn't provide the fund: they had to prioritise and feed the fund somewhere else.

If the initiative did not achieve the intended outcomes, what do you believe were the reasons for this?

Lack of funding and support from policy makers.

Based on your learning from this initiative, what recommendations would you make for future best practice guides?

More support from policy makers, and better publicity plan for our solutions.

Case study 4: Monitoring attitudes towards immigrants

Overview

This is an ongoing project which aims to measure and monitor changes in attitudes towards immigrants during the early stages of the current COVID-19 outbreak in five countries: Germany, Italy, Spain, the United Kingdom and the United States, using Twitter data and natural language processing.

Who funded this initiative?

The United Nations's International Organization for Migration.

Description / intended outcomes

The project seeks to determine the extent of intensification in anti-immigration sentiment as the geographical spread and fatality rate of COVID-19 increases; identify key discrimination and racism topics associated with anti-immigration sentiment; and assess how these topics and immigration sentiment change over time and vary by country.

Overall, did the initiative make the desired impact?

Not known at this stage.

What worked well?

Close collaboration with key stakeholders; development of a framework leveraging on new forms of data (Twitter) to measure immigration sentiment in a scenario in which traditional data sources were unavailable; provide novel insights in near real-time.

What were the difficulties and challenges?

Accessing and processing the data.

If the initiative did not achieve the intended outcomes, what do you believe were the reasons for this?

The complexities of extracting meaningful information from tweet posts.

Based on your learning from this initiative, what recommendations would you make for future best practice guides?

Develop a comprehensive understanding of the use of Twitter APIs. They are expensive and requests need to be carefully planned. A carefully designed list of search terms is key as well as knowing that Twitter retrieves information from a point backwards.

Case study 5: Predicting the extent of COVID-19 impact

Organisation: De Montfort University.

Overview

We used simple, publicly available data to predict the extent of COVID-19 impact in the 632 parliamentary constituencies of Great Britain. We also proposed and modelled a scenario that showed that up to 25% of GB could have avoided a lockdown.

Who funded this initiative?

Self-funded.

Description / intended outcomes

The initiative was to use simple, publicly available data to model the spread and impact of COVID-19. This could then be used to "protect" certain areas from spread whilst treating infected areas.

Overall, did the initiative make the desired impact?

Not known at this stage.

What worked well?

1. Gathering simple, disparate, publicly available data.
2. The AI modelling using self-organising maps.
3. The accuracy of the modelling as the pandemic progressed.

What were the difficulties and challenges?

There were no challenges as such, it is just a case of better (more accurate) data would result in more accurate modelling. Getting more accurate data requires government involvement.

If the initiative did not achieve the intended outcomes, what do you believe were the reasons for this?

The ideal outcome would have been for the government to pilot this in a few regions. In essence we argue that they did pilot it (e.g. preventing people from travelling from Tier 3 to Tier 2 etc.), but not to the extent that we would have liked.

Based on your learning from this initiative, what recommendations would you make for future best practice guides?

AI and data analytics can certainly make a rapid and significant positive impact to pandemic modelling and containment, and governments should focus more attention in this area.