Data Circulation in Health Landscapes

Niccolò Tempini
University of Exeter

Antonio Maturo
University of Bologna

Elisabetta Tola
Formicablu

Abstract: The crossing boundaries intends to open a dialogue between Science and Technology Studies, Social studies of Health and the emerging Data Journalism perspective. It explores major issues at stake in contemporary practices of producing and sharing data, with a focus on the COVID-19 pandemic.

Keywords: health data; platforms; risks; pandemic; data journalism.

Submitted: April 4, 2022 – **Accepted:** June 10, 2022

Corresponding author: Niccolò Tempini, Department of Sociology, Philosophy and Anthropology, University of Exeter & Alan Turing Institute, Byrne House FF8, Exeter (UK), n.tempini@exeter.ac.uk

Pandemic Data Circulation

Niccolò Tempini, University of Exeter

Introduction

We have been two years in the COVID-19 (C19) pandemic, and many interesting patterns have emerged that are worth discussing. I can only attempt to touch on a few of them, which are related to the practices and flows of health data. It was interesting to see, as the pandemic ensued, how many different kinds of data were mobilised. And many different social actors got involved in the use of data, for many different purposes.

Data were put in circulation in ways and speed that were unforeseen, from both public and private sector companies. But data circulate well in some directions, and less so in others. Overwhelmed perusing a constantly moving panoply of numbers, charts and assessments on the state of the pandemic, it is easy to miss out that some data are not flowing well at all, and that others should perhaps stay where they are.

UK: "State of Play"

The UK has been widely regarded as one country where the state has been most willing to play with various sorts of data experimentation. A report by the Alan Turing Institute on data science and artificial intelligence in the "age" of COVID-19 (von Borzyskowski et al. 2021) highlights how, in many ways, the pandemic was an exceptionally propitious opportunity for all sorts of innovation and experimentation with data to occur. As it started reacting to the C19 outbreak, on 17th March 2020 the Government mandated healthcare data custodians at NHS (National Health Service) Digital to support access and processing of health data by authorised organisations for purposes of pandemic response (Secretary of State for Health and Social Care 2020). The relaxation of regulatory standards, with a mandate for various public agencies to put the data in motion, was the first "booster" for the circulation of data. But it was not only national healthcare system data that were quickly mobilised. Private companies including the likes of Silicon Valley giants Google and Apple also offered select access to proprietary data, albeit to a limited extent. This created the space for various teams and organisations to intervene and offer their services as to how such an unprecedented and all-encompassing mobilisation could be achieved. Datasets that could matter for the coordination of response were myriad, as were the indicators to closely watch to monitor how things unfold. Government decision makers procured from private companies a set of analytics computational infrastructures to manage the former, and analytics dashboards for watching the latter. One of the most important contracts, worth more than 12.5 million GBP (Gov. UK 2020), was awarded to Palantir – a secretive Silicon Valley company familiar to controversies thanks to its propensity sourcing contracts from military, intelligence and border control agencies involved in politically questionable missions. Consultancies McKinsey, Deloitte and Faculty AI also contributed. On top of Palantir's computational infrastructure NHS Digital could then launch a "COVID-19 Data Store", a repository of datasets available to agencies involved in the pandemic response: *"NHS COVID-19 Data Store brings together and protects accurate, real-time information to inform strategic and operational decisions in response to the current pandemic in one place."* (NHS England 2022a) The datasets included in the Data Store are rather disparate and come in many different formats, as a collection which could be potentially accessed by many different users for many different purposes would be (NHS England 2022b). There are data such as raw NHS medical helpline call data; NHS staff absence reports; mobility data from Google and Apple; Enterprise Resource Planning data for healthcare system management; patient demographics; counts of online and video consultations; personal protective equipment stocks and purchases; and even self-report symptom data collected by volunteers for the COVID-19 Symptom Study using a citizen science app called "ZOE". These data are

not limited to tables of numerical and/or structured text values. They are also lists, lookup tables, summaries of historical data, and records that "describe".

Thanks to several dashboards and analytics features which draw disparate data sources together in "actionable" summaries of visualisations, charts and numbers, decision makers in government should be equipped with the best real time intelligence to take complex decisions: *"These dashboards are designed to help senior national and regional officials to make policy and strategic decisions in response to COVID-19"* (NHS England 2022a). Perusing Palantir's contract we learn that the system comprises of three main interfaces: the Strategic Decision Makers Dashboard; Recovery of Critical Services; and Early Warning System. We also learn that these technologies might have a scope and longevity that reach farer than the pandemic alone: they should help to *"coordinate national response to COVID-19 and EU Exit"* and provisions are made to allow the Government to transition this system from the pandemic use to *"general business-as-usual monitoring"* (if Palantir's software-as-a-service contract is renewed past expiry). The Early Warning System interface seems the most ambitious, sporting an *"Explainability and Trust Overview"* feature displaying forecasts generated by the models of a private third party (the consultancy Faculty AI, founded by a physicist) using NHS data, 111 medical helpline data, and Google and Apple mobility data among others. And so, the time when governing comes to resemble a session of Sim City (or Chile's Cyber-Syn room of cybernetic government, discussed by Medina 2015) might be finally here. Those who sit at the fence of government action and have limited information to go by might have many questions about this "system of systems". One may wonder, for instance, who is interested in knowing counts of tele-care consultations? How many of the people calling a medical helpline or logging their symptoms through the ZOE app would imagine that their data could show up on a government dashboard, and what would they think if they knew? Are self-report data from the ZOE COVID Symptom Study app displayed on any dashboard, and who looks at it?

Those around the Data Store are not the only movements of health data between public and private sector that are currently noteworthy in the UK. The General Practice Data for Planning and Research (n.b. GPDPR – not GDPR) is a policy unveiled in the Spring of 2021 that, resurrecting the ambitions of defunct "*care.data* framework" (Vezyridis and Timmons 2017), mandates NHS Digital to create a centralised repository of general practitioner healthcare data, the access to which should be sold to private sector companies on a cost-recovery basis. It is the latest of a series of attempts to enable the permanent circulation of national healthcare system data in the UK private sector and boost its valorisation. As researchers at the Ada Lovelace Institute observe (Machirori and Patel 2021), the scheme was introduced with notable disregard for public engagement through a "method" that could be described as "decide, announce and defend": if

the policy is rammed through fast enough, it might survive the public backlash and the government would get its way. Once again, public backlash might prove sticky enough. The introduction of the plan has been delayed and there has been a sizeable opt-out, as noted by Cori Crider, director of Foxglove Legal, in a recent expert consultation by the Ada Lovelace Institute (2021). Regardless the exact timings, the repeated attempts of consecutive UK governments to enable private sector use of patient records demonstrate a long-term determination, which predates and will outlast the pandemic, to get health data to circulate more widely and loosely and for many more purposes than the performance of health care; and for the government to allow private data platform developers to embed themselves in the infrastructure of the state and its governance activity. As also highlighted by the Ada Lovelace Institute, the UK Government fancies the opportunity to turn the relaxation of data circulation regulations introduced during the pandemic into a standard for the future regulatory regime, so as to favour faster and broader circulation. The same kind of pattern has been observed in yet another front of government development of pandemic technology infrastructure, that of contact-tracing apps. As Rob Kitchin has noted (Kitchin 2020), in order to develop contact-tracing apps many countries desperate to curb the spread of C19 resorted to working with organisations that have been at the centre of scandals or polemics in "normal" time because they develop controversial techniques, technologies or services of population surveillance. Contact-tracing collaborations includes organisations such as NSO Group, who sells weapons-grade spyware to illiberal and autocratic governments accused of repressing dissidents and opponents, and has worked with Israel on their app. More notoriously, tech giants and mobile monopolists Google and Apple, who rushed to offer a common Bluetooth-based stack for automating contact tracing in Android and iOS phones, used their privileged position of mobile gatekeepers to make an impactful contribution (not without privacy implications – see Kitchin 2020).

Translating Private Technology to Public Infrastructure

In respect to organisational operations and decisions, data seem to circulate well indeed. Many more private organisations have been taking part in the C19 data craze, often with much display aimed at "covid-washing" their reputation (Kitchin 2020), keen to be seen as generous tech wizards rather than greedy data harvesters. After all, one might say, there is a point in letting these companies collecting so much data about the public, if they respond to the call when their help is needed. But are in particular those organisations, who outside of "pandemic time" have been at the centre of many ethical controversies over the ways in which they generate and use data, that have rushed to the forefront of more and less consequential

efforts to help. It reveals a key assumption as to the ways in which technological, organisational and methodological frameworks originally developed for watching and manipulating consumer behaviour through digital technology have been seen as *translatable* to the context of social distancing restrictions and other emergency rules. The double-edgedness of these initiatives is easy to surmise. For instance, data broker Experian, who sells individual data at population scale after collecting them through a vast network of business relationships and repackaging them in the form of value-added marketing demographics, studied the distribution of C19's socio-economic impacts. Besides the potential to help public health response, one should remember the knowledge thus generated is likely to have value for marketing demographics too; and so, the first beneficiary of this effort might well be Experian itself. At least in a rhetorical sense, these frameworks have proved translatable: research into individual attitudes towards contact-tracing apps (Lucivero et al. 2021) shows that many believe the impact of intensive data collection and cross-dataset linking is negligible since the lives of ordinary individuals are already intensively surveilled, and for much less of a reason.

With so much "help" on offer, the pandemic has certainly reaffirmed the central role of private technology in the coordination of society's reaction to emerging events. But should it? A piece on the Harvard Business Review (Balsari et al. 2020) suggests otherwise. A *"tidal wave of data"* is sloshing around all corners, but not much of it might be *"any good"*. Many datasets made available are incomplete in ways that are not-randomly distributed across society, but rather, reflect socio-economic inequalities. If the disadvantaged are less well represented in datasets used to coordinate pandemic response, expect the inequality to be drawn on. And so, the authors suggest, while many tech organisations are busy offering up datasets and expertise on linkage, hosting and analytics, there is not enough engagement with subject matter experts. Many models are produced with expertise that is translated from being involved in the solution of problems other than the medical, but rather, rooted in mathematics, physics, or operations management expertise, among others; other innovations, such as automated contact-tracing, are live experiments. What a contrast with the exhortations of data analytics and visualisations leader Tableau, who encourages users to start *"your own analysis"* (Tableau 2022). Cloud-computing giant, Amazon Web Services, offers a suite of data and computational infrastructure resources to help and *"provide these experts with the data and tools needed to better understand, track, plan for, and eventually contain and neutralize the virus that causes COVID-19"* (Amazon Web Services 2020).

Experts can *"use AWS or third-party tools to perform trend analysis, do keyword search, perform question/answer analysis, build and run machine learning models, or run custom analyses to meet their specific needs."*

The Challenges that Remain

While some data might have been circulated in and out, and across, government quite well, other data were not circulating equally well. As the Alan Turing Institute's report observes (von Borzyskowski et al. 2021), a number of challenges were experienced by the community of data science and artificial intelligence researchers striving to make an impact through the production of new knowledge about C19. Certain kinds of data can be difficult to access because of governance issues – the Ada Lovelace Institute points out that current governance processes were often too slow and required too much of too few data custodians (Ada Lovelace Institute 2021); but also because they are more difficult to generate than others. Data on some pandemic response measures and their impacts, such as non-pharmaceutical interventions (e.g., social distancing and face masks), were not sufficiently available. Local council and administration decision makers complained not enough data were made available to them. Uneven quality and representation in population datasets further raised concerns of inequality in the response to the pandemic. Unequal vulnerability to pandemic response measures would also lead to mistrust and uneven participation and compliance in various undertakings, such as active installs of contact-tracing apps, or symptom self-reporting in citizen science studies such as ZOE COVID Symptom Study. While datasets were over-produced in a scramble to help, attention slipped over quality and methodological issues such as sampling (von Borzyskowski et al. 2021). While new problems are often resolved more quickly the more open and participative is the search for a solution, there is a way in which the eventually ensuing chaos brings about new problems in the meantime. The ubiquitous discussion of statistics and data in all kinds of public reporting further amplified concerns over interpretation and communication. Last but not least, Alan Turing Institute researchers complain about their relationship with government decision makers. They found it difficult to understand if the expert knowledge that they were generating through many studies was getting any attention by policy makers. Researchers who are well connected could have government's ear and access data that others could not. As we have seen, government decision-makers were providing themselves with cutting-edge analytics technology from private firms the likes of Palantir. It is as if they wanted to lock themselves up in the button room with the latest tech gear,

leaving other experts outside who kept insisting they could help. The appeal of translating all-powerful consumer surveillance infrastructure might have been more powerful than working with experts the old way.

From Normal to Pandemic Time, and Back

From this quick sketch, it should be possible to get a feel for a complex and protracted situation involving many kinds of initiatives, data practices, actions, claims, and contexts of use. Talk about data can be at different levels (Rosenberg 2013; Leonelli 2016; Tempini 2020): as digital object stored on computer systems, as epistemic product of an empirical scientific investigation or of activist projects, and as a rhetorical device that can be waved around (as in the press conferences that saw Boris Johnson so frequently argue that the UK government "just" followed the data). This makes the analysis of data practices and movements complex. Some of the stakes of data practices and movements from "pandemic time" will be played out in the future, and in such a cacophony of data practices and claims it is easy to lose sight of a few big trends that have been driving all things data. But there is perhaps enough to see that the circulation of data during the pandemic was uneven and dependent on many factors including the organisational and technological infrastructures datasets are managed with; and to suspect that the pandemic crisis, like many other crises, won't be let go to waste, and instead, will allow private infrastructure to be wedged further underneath society and its spaces. In times of emergency a general mobilisation of all sectors and actors might feel like the only intelligent thing to do, and so many private sector organisations all scrambled to see what they could offer to government and research community. But in all things infrastructural, there is a strong sense in which the past will become the present, and the present will become the future; because infrastructures are built over long time and sit on top of even longer-evolving methodological and cultural frames, path dependencies are deemed and continued through the material shaping of systems and practices (Star and Ruhleder 1996; Hanseth, Monteiro, and Hatling 1996; Hanseth 1996; Bowker and Star 1999).

Time was of the essence, and the rapidity with which private data-intensive businesses deployed a panoply of initiatives to share and analyse data of all kinds bears witness to the strategic dynamism and translatability of data platforms. Response to the pandemic was characterised by lack of time, and those who control data infrastructures were in a good position to enter the frame of pandemic response efforts and reap financial and

reputational benefits. Infrastructure evolution in "normal" time has often been seen as challenged by "infrastructural inertia" (Star and Ruhleder 1996; Bowker and Star 1999), that is the way in which infrastructures swamp change and make material legacy. In "pandemic time", instead, it seems as if infrastructures were key to enable movement and change. All time had been suddenly sucked out. Being able to re-deploy and re-purpose data infrastructures and methods was a chief way to buy time. This might not need to be a contradiction. Infrastructural inertia is likely to be observed as an infrastructural reaction when the change that is being carried out is proactive or transformative – a move away from the current ways of doing. The change and re-organisation that the pandemic time required was essentially *reactive*: when surprised and unprepared to unexpected developments, the current ways of doing might be the only available to effect change. Infrastructures and methods that are re-deployable, transferable and extendable are quickly whisked into new positions.

One could wonder why should we be concerned about all this? That is because once time is "normal" again, infrastructural inertia can kick in again. Dislodging private technology infrastructure, and the practices associated with it, that was deeply embedded in the government machinery back in pandemic time will become ever more difficult. As Sharon points out (Sharon 2020), running pandemic response with the technology provided by private corporations will *"increase our dependency on them for the provision of (public) services, and they make themselves necessary passage points for the adequate functioning of these sectors"* in the future. Continued reliance on any infrastructure makes it invisible and undermines the imagination of technological-organisational-political alignments that respond to different values, priorities and logics. Complex and consequential infrastructures, and their developmental inertia, help to ensure the past, from before the pandemic, is carried over to what comes after. They make some of the linkages that thread together the before-during-after of pandemic times.

Update 9th June 2022

This morning the Financial Times is breaking with reporting suggesting the Government is planning to award a giant contract for the provision of a data analytics "operating system for the NHS" and Palantir is devoting enormous resources to win the contract. Privacy activists who have exposed Palantir's penetration in state infrastructure since 2021 point out the same kinds of concerns I have been repeating here. It turns out worrying developments might be moving even faster than we might have worried.

"Palantir gears up to expand its reach into UK's NHS", 2022. Financial Times. https://on.ft.com/3xaqnsw

References

Ada Lovelace Institute (2021) *Learning Data Lessons: Data Access and Sharing during COVID-19*, London, Ada Lovelace Institute & the Royal Society.

Amazon Web Services (2020) *A Public Data Lake for Analysis of COVID-19,* in https://aws.amazon.com/blogs/big-data/a-public-data-lake-for-analysis-of-covid-19-data/ (retrieved June 3, 2022).

Balsari, S., Buckee, C. and Khanna, T. (2020) *Which Covid-19 Data Can You Trust?* in "Harvard Business Review", in https://hbr.org/2020/05/which-covid-19-data-can-you-trust (retrieved June 3, 2022).

Bowker, G.C. and Starr, S.L. (1999) *Sorting Things out: Classification and Its Consequences*, Cambridge, MIT Press.

Gov. UK (2020) *G-Cloud 11 Call-Off Contract*, in https://www.contractsfinder.service.gov.uk/Notice/Attachment/0e1e4c85-6d57-4fca-9cc8-88adcffb88a2 (retrieved June 3, 2022).

Hanseth, O. (1996) *Information Technology as Infrastructure*, Doctoral Dissertation, Göteborg, Göteborg University.

Hanseth, O., Monteiro, E. and Hatling, M. (1996) *Developing Information Infrastructure: The Tension Between Standardization and Flexibility*, in "Science, Technology & Human Values", 21 (4), pp. 407-426.

Kitchin, R. (2020) *Civil Liberties or Public Health, or Civil Liberties and Public Health? Using Surveillance Technologies to Tackle the Spread of COVID-19*, in "Space and Polity" 24 (3), pp. 362-381.

Leonelli, S. (2016) *Data-Centric Biology: A Philosophical Study*, Chicago, University of Chicago Press.

Lucivero, F., Marelli, L., Hangel, N., Zimmermann, B.M., Prainsack, B., Galasso, I., Horn, R., Kieslich, K., Lanzing, M., Lievevrouw, E., Ongolly, F., Samuel, G., Sharon, T., Lotje, S., Stendahl, E. and Hoyweghen van I., (2021) *Normative Positions towards COVID-19 Contact-Tracing Apps: Findings from a Large-Scale Qualitative Study in Nine European Countries,* in "Critical Public Health", 32 (1), pp. 5-18.

Machirori, M. and Patel, R. (2021) *Turning Distrust in Data Sharing into Engage, Deliberate, Decide*, in https://www.adalovelaceinstitute.org/blog/distrust-data-sharing-engage-deliberate-decide/ (retrieved June 3, 2022).

Medina, E. (2015) *Rethinking Algorithmic Regulation*, in "Kybernetes", 44 (6/7), pp. 1005-1019.

NHS England (2022a) *NHS COVID-19 Data Store*, in https://www.england.nhs.uk/contact-us/privacy-notice/how-we-use-your-information/covid-19-response/nhs-covid-19-data-store/ (retrieved April 29, 2022).

NHS England (2022b) *COVID-19 Datastore Reference Library*, in https://data.england.nhs.uk/covid-19/ (retrieved April 29, 2022)

Rosenberg, D. (2013) *Data before the Fact*, in L. Gitelman (ed.), *Review Data Is an Oxymoron*, Cambridge, MIT Press, pp. 15-40.

Secretary of State for Health and Social Care (2020) *Control of Patient Information (COPI) Notice*, in https://digital.nhs.uk/coronavirus/coronavirus-covid-19-response-information-governance-hub/control-of-patient-information-copi-notice (retrieved April 29, 2022).

Sharon, T. (2020) *When Google and Apple Get Privacy Right, Is There Still Something Wrong?*, in https://medium.com/@TamarSharon/when-google-and-apple-get-privacy-right-is-there-still-something-wrong-a7be4166c295 (retrieved June 3, 2022).

Star, S.L. and Ruhleder, K. (1996) *Steps Toward an Ecology of Infrastructure: Design and Access for Large Information Spaces*, in "Information Systems Research", 7 (1), pp. 111-134.

Tableau (2022) *COVID-19 (Coronavirus) Data Hub*, in https://www.tableau.com/covid-19-coronavirus-data-resources (retrieved April 29, 2022).

Tempini, N. (2020) *The Reuse of Digital Computer Data: Transformation, Recombination and Generation of Data Mixes in Big Data Science*, in S. Leonelli and N. Tempini (eds.), *Data Journeys in the Sciences*, Cham, Springer International Publishing, pp. 239-263.

Vezyridis, P. and Timmons, S. (2017) *Understanding the Care.Data Conundrum: New Information Flows for Economic Growth*, in "Big Data & Society" 4 (1), doi:10.1177/2053951716688490.

von Borzyskowski, I., Mateen, B., Mazumder, A. and Wooldridge, M. (2021) *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*, London, Alan Turing Institute.

* * *

Polysocial Risk Scores and Behavior-Based Health Insurance: Promises and Perils

Antonio Maturo, University of Bologna

Cotton Balls, Zinc Supplements and Predictive Analytics

Once upon a time, a long time ago, around 2010, an irate father walked into a Target store on the outskirts of Minneapolis. He asked to speak with the manager, and upon their arrival, he waved coupons and vouchers in their face:

> "My daughter got this in the mail!" he said. "She's still in high school, and you're sending her coupons for baby clothes and cribs? Are you trying to encourage her to get pregnant?"[1]

The manager apologized profusely and stammered that he had no idea how this could have happened.

A few days later, the same manager called the father to apologize again, but something happened:

> On the phone, though, the father was somewhat abashed. "I had a talk with my daughter," he said. "It turns out there's been some activities in my house I haven't been completely aware of. She's due in August. I owe you an apology." (ibid.)

What led to this bewildering encounter was a new office in that retail location, where a mysterious new practice had been implemented: *Predictive Analytics*. A sudden change in the young woman's shopping patterns had been noticed, signaled through her loyalty card, sparking an unanticipated chain reaction. Back in 2010, retailers had just started to collect intimate details about consumption habits. They had noted that:

> Women on the baby registry were buying larger quantities of unscented lotion around the beginning of their second trimester. Another analyst noted that sometime in the first 20 weeks, pregnant women loaded up on supplements like calcium, magnesium and zinc. Many shoppers purchase soap and cotton balls, but when someone suddenly starts buying lots of scent-free soap and extra-big bags of cotton balls, in addition to hand sanitizers and washcloths, it signals they could be getting close to their delivery date.[2]

Because of this shift in purchasing habits, the young woman's pregnancy had been made apparent in her data-double, even before her social identity.

This incident occurred over ten years ago, while paper mail was still the main form of promotion. In the meantime, self-tracking has exploded, generating enormous amounts of data, especially physiological and behavioral data. In addition, sophisticated algorithms can monitor the time we spend on a site, the physical places we visit, and the likes we place. By monitoring our credit cards, it is possible to know what we eat and how many calories we ingest. Especially in the context of COVID-19, unseen sensors can recognize who is running a temperature in a train station. In the field of health, therefore, there is not only big data but thick data: data that can tell us about our health from a clinical, physical and social point of view.

In this datafication of health, perhaps the two most disruptive and cutting-edge developments are "Polysocial Risk Scores" and "Behavior-Based Health Insurance". These two areas, in some ways, overlap since the risk score is the basis of health insurance, of which Behavior-Based is the most advanced kind. Surrounding both are big players with keen interests and high expectations. Both Polysocial Risk Scores and Behavior-Based Health Insurance share close attention to social aspects of health, and both are

driven by the need to predict possible (individual) futures on the basis, of course, of quantification (e.g., datafication). Surrounding these developments, beyond innovative possibilities, are clear doubts and concerns about their implications and consequences in terms of social justice.

The Polysocial Risk Score

In order to understand what Polysocial Risk Scores consist of, it is helpful to underline the main features of the Polygenic Risk Score, which in some ways acts as its prototype.

The Polygenic Risk Score estimates the risk that a person has of developing a disease from his or her genes. More precisely, the Polygenic Risk Score represents the total number of genetic variants that an individual has to assess their heritable risk of developing a particular disease since multiple genetic mutations and their interactions cause most diseases.

At first glance, Polysocial Risk Scores can be seen as the sociological version of the Polygenic Risk Scores, with the idea of the Polysocial Risk Score being developed in the context of the social theory of social determinants of health.

Social determinants of health are the factors that affect a person's health, namely education, income, type of work, type of housing, neighborhood, social cohesion, and others. These determinants affect health through lifestyles, health literacy, and access to care. Epidemiologists and health sociologists have repeatedly confirmed the influence of social context and social determinants on physiology. The determinants of health are strongly intertwined, e.g., how income influences health and how it, in turn, is affected by education; how the weight of income and how the weight of education affects a person's health, and how much, in turn, the weight of education on the possibility of acquiring higher income.

Therefore, the challenge is to weigh and estimate the conditioning of social determinants and their interactions with individual health. However, to date:

> Most efforts to precisely quantify the influence of individual social determinants of health have failed, largely because the causal pathways are numerous, interconnected, and complex. (Figueroa et al. 2020, 1553).

The enormous amount of data that can now be acquired on people's health could mark a turning point for developing precise estimates of individual risk of becoming ill. Notwithstanding, one would have to arrive at a Polysocial Risk Score for each disease or health outcome, even in this case. One person would then have several Polysocial Risk Scores. Nevertheless, compared to the Polygenic Risk Score, there is a considerably more turbulent level of complexity:

> One key difference is that unlike polygenic risk scores, which are not dynamic because the scores are based on an individual's genes, polysocial risk scores may change if an individual's social circumstances change. (ibid.)

Where the Polygenic Risk Score is static, the Polysocial Risk Score is (would be) dynamic. Moreover, the same social determinants have different weights in different social contexts. Indeed, and methodologically it is even worse with some social determinants being part of the context itself (e.g., social capital and social cohesion).

As Figueroa and collogues (2020) illustrate, it is necessary to constantly collect, aggregate, and mobilize data from different domains regarding people's quality of life and sociodemographic data Polysocial Risk Scores need to be periodically updated. Above all, it is necessary to relate these "external" data to people's state of health, to their "internal" health data, and to their physiology.

Moreover, as scores are elaborated and processed by algorithms, in some cases, health data may result in biases and, in worst cases, social discrimination. As summarized by Leslie et al.:

> AI systems can introduce or reflect bias and discrimination in three ways: in patterns of health discrimination that become entrenched in datasets, in data representativeness, and in human choices made during the design, development, and deployment of these systems (2021, 1).

Thus was the case of genetic data, as in the U.S, most genome-wide association study-based polygenic risk scores have been based on populations of European descent, neglecting the health of other ethnic minorities.

Pricing Risk: Behavior-Based Health Insurance

Creating the Polysocial Risk Score would be something between miracle and mirage, yet this does not mean that attempts have not been made. On the contrary, the health analytics industry is a rapidly developing sector in the digital firms of Silicon Valley and the biotech industry of the Boston Area, with the American health insurance agencies leading the charge towards the construction of health risk scores, with the latter being interested in knowing the health status of their members. Moreover, actors that has most influenced this orientation of health insurance, at least according to some scholars, has been a legal provision contained in the Affordable Care Act (ACA), approved in 2010. As Liz McFall points out:

> The ACA alternative introduced a "behavioural" approach (...) including new responsibilities to pay a "fair share" of the costs of the entire pool and be "as healthy as you can." The responsibility to be healthy is promoted by the provision of access to preventative care and treatments for chronic,

> preventable disease. (...) This emphasis on behavioral responsibility is a great fit with data-driven healthcare innovations including wearable self-tracking devices and apps. (Mc Fall 2019, 60).

This provision has operated in "association" with other factors, primarily technology. As McFall (2019) and Schüll (2016) point out, digital technology and the ACA have been presented as a "dynamic duo" working together, and

> compelling insurers, health care providers and consumers to cut costs (...) shifting the management of chronic conditions like diabetes and heart disease away from hospitals and doctors and into the hands of patients themselves (Schüll 2016, 318).

If over a decade ago the office of a chain store was able to learn of a customer's pregnancy through her purchases of hygiene products, what can health insurers know about us today? What could insurance "providers" learn when they are given access to sociodemographic data, clinical data, genetic predisposition, and, more importantly, lifestyle data (not simply "lifestyle data" as in whether individuals are smokers or vegetarians, but all digital activities and data-doubles)? Moreover, some digital platforms have already identified rich sets of data points for proxies of social determinants of health:

> individual purchasing behavior, consumer engagement with advertising, insurance claims, sentiment, and expression in online forums, credit histories, and online social networks (Rowe 2021, 4).

This data, in turn, is coupled with the mundane data generated by personal FitBits, generously gifted by health insurance agencies (Maturo and Moretti, 2018).

Before the spread of digital social networks, Christakis and Fowler (2010) wrote that social friend networks greatly influence personal decisions. Christakis and Fowler showed through animated sociograms based on accurate longitudinal research how certain behaviors may be "contagious". Not only does a person have a high probability of gaining weight if their friend does, but also if their friend's friend does, this can be further applied to divorce and smoking cessation. Today these analyses are immensely easier given the ease with which big data can be collected and processed. The predictive potential delivered to insurance agencies is enormous, leading to correlation taking the place of causation, with the latter becoming an obsolete 20th-century category (Anderson 2008).

Raschel Rowe (2021) has done thorough research on the platform "Opioid360", a platform that combines browser histories, credit, insurance, social media, and traditional survey data to sell the service of risk

calculation in population health. Created as a tool that would support overworked clinicians to see invisible signs of potential addiction in their patients, Opioid360 paved the way for broader applications to prevent chronic diseases. Most importantly:

> By extending digital phenotyping imaginaries, Opioid360's presentation appealed to the notion that comprehensive personal data can offer behavioral science the precision that genomics has offered to identify rare diseases (Rowe, 2021, 4).

In their analysis of Vitality health insurance, McFall et al. (2020) make clear that:

> Behaviour is Vitality's core brand value and its policies provide incentives to customers to meet behavioral targets, share their data with the company and share their progress on social media (McFall et al. 2020, 7).

The big switch that many health insurers have made is to link insurance premiums and access to specific policies to the constant digital monitoring of physical activity (InsurTech). In theory, through self-tracking, the premium costs could fluctuate every day, in connection with our physical states, instead of once a year. The extension of insurance surveillance to other aspects of our lives through the datafication of health raises big questions about social justice.

The encouragement of certain behaviors opens an extended reflection on the empowerment of the individual. In social studies of health, it is well known how social context affects a person's health and that certain social factors such as income make adherence to healthy lifestyles relatively easy for some people, while for others practically impossible.

When I arrive at around 8 o'clock outside my department, I often meet one of the ladies who clean the offices – being female, visibly overweight, doing an extremely physical job (maybe she has a disease or seeks satisfaction in food?). She gets up at 4.45 a.m. to start work before 6 a.m. When she greets me at 8 a.m., she lights a cigarette with her South Italian accent before getting into the car. She inhales in big puffs as if it were a prize, a seal, or as we say today in the field of gamification, an award for the work done. However, it is not her avatar who is smoking, unfortunately. Her face is tired, and she is in a hurry – maybe she will light another one soon: she has to go to the other side of the town to do some more cleaning, and there is a lot of traffic by then. Just before entering the department, out of the corner of my eye, I see a colleague of mine jogging through the beautiful palm trees on our campus.

Algorithmic Forecasting and Insurance Customization

According to Barry and Carpentier (2020), insurance can be defined as

> the transformation of unknown individual uncertainty, or chance, into a measurable aggregate risk. Technically, it consists of pooling uncertainty and applying the law of large numbers (Barry and Carpentier 2020, 3).

In this way, the occurrence of catastrophic events for one person was remedied by adding small amounts set aside by all. Through statistical predictions, it is relatively easy to predict that a certain number of insured people will fall ill without knowing who exactly. At least until now, insurance has been based on the concept of socialized actuarialism. However, as early as 1996, O'Malley glimpsed the advance of privatized actuarialism, a more refined approach based on:

> a technology of governance that removes the key concept of regulating individuals through collectivistic risk management and places the responsibility for risk management back on the individual (O'Malley 1996, 197).

Thus, whereas traditional insurance was based on prediction (i.e., aggregate predictions at the macro level), the new behavior-based insurance is based on forecasting (i.e., attention to the individual's future at the micro-level). This mode of insurance makes policyholders more responsible for their daily actions and health. However, many scholars question whether, technically, behavior-based insurance can still be considered insurance. Based on the distinction between individual fairness and social fairness, Cevolini and Esposito, effectively summarize how the ancient principle of solidarity can be undermined by new insurance policies:

> Algorithmic prediction could radicalize the principle of segmentation, culminating in the extreme case of "segments of one." This would almost automatically mean the end of the risk-pooling on which the principle of risk-sharing is based (Cevolini and Esposito 2020, 4).

The end of risk-pooling carries significant implications as to whether Polysocial Risk Scores have the potential to become a central tool in healthcare. In this regard, a crucial issue here concerns what would happen if Polysocial Risk Scores are calculated and accredited by institutions.

Considering that constructed indicators tend to become objective entities, Polysocial Risk Scores can be employed in different contexts and by different actors; from public health departments, government officials,

technology companies, investors, and private insurance companies (Neresini 2015). In a world that is increasingly computerized, quantified, and managed by algorithms, health scores could be mobilized for a variety of purposes. Some of these uses could be noble and others less so:

> Health risk scores are not only useful for immediate patient classification or public health program planning, they are also useful to investors seeking to leverage or hedge their risk exposure. (Rowe 2021, 9).

Although indirectly, a strong impetus for developing health scores has undoubtedly come from COVID-19 pandemic. The pandemic has bolstered the trend of health quantification through the robust joint growth of medicalisation and digitalisation. Most importantly, COVID-19 pandemic has spurred surveillance. To put a long story short: *9/11 increased police surveillance, big data stimulated capitalist surveillance, and COVID-19 hyperbolically accelerated molecular surveillance*. Molecular surveillance can be seen as the scrupulous and precise monitoring of our physiological motions and their instantaneous transformation into data. A panopticon of our internal states, or more precisely: the *endopticon* (Maturo 2015). However, this surveilling is not perpetrated by shadowy officials of mysterious agencies wearing thick-lensed glasses in smoke-filled rooms of some governmental molecular surveillance departments but by algorithms themselves. Programs that react to numbers that exceed certain thresholds, to parameters that measure, compare, and discriminate our physiological motions, collect our behavioral habits and read our molecules' silent but vivacious lives.

Yuval Noal Harari, the author of the successful *Homo Deus*, in an article published in the *Financial Times* on April 19, 2020 entitled *The world after the Coronavirus*, fears a dystopian scenario:

> Hitherto, when your finger touched the screen of your smartphone and clicked on a link, the government wanted to know what exactly your finger was clicking on. But with Coronavirus, the focus of interest shifts. Now the government wants to know the temperature of your finger and the blood-pressure under its skin. One of the problems we face in working out where we stand on surveillance is that none of us know exactly how we are being surveilled, and what the coming years might bring. Surveillance technology is developing at breakneck speed, and what seemed science-fiction 10 years ago is today old news.[3]

Harari's concerns reaffirm that health scores will soon be the subject of a Black Mirror episode. Behavior-Based Insurance and Polysocial Risk

Score have disturbing implications, starting with the de-politicization of health, which is no longer understood as a public and social issue but as a business and private concern. The challenge, however, is not to assume ipso facto Luddite or apocalyptic attitudes. It is necessary to find a catalyst that brings health back to the center of public discourse. In a society dominated by chronicity, the masses (of patients and caregivers) should become aware of their strength.

References

Anderson, C. (2008) *The end of theory: The data deluge makes the scientific method obsolete*, in https://www.wired.com/2008/06/pb-theory/ (retrieved June 3, 2022).

Barry, L. and Charpentier, A. (2020) *Personalization as a promise: Can Big Data change the practice of insurance?*, in "Big Data & Society", 7 (1), pp. 1-12.

Cevolini, A. and Esposito, E. (2020) *From pool to profile: Social consequences of algorithmic prediction in insurance*, in "Big Data & Society", 7 (2), pp. 1-11.

Christakis, N. and Fowler, J. (2011) *Connected: The Surprising Power of Our Social Networks and How They Shape Our Life*, Boston, Little Brown and Co.

Figueroa, J.F., Frakt, A.B. and Jha, A. (2020) *Addressing Social Determinants of Health: Time for a Polysocial Risk Score*, in "Journal of American Medical Association", 323 (16), pp. 1553-1554.

Fourcade, M. and Healy, K. (2013) *Classification Situations: Life-chances in the Neoliberal Era*, in "Accounting, Organizations, and Society", 38 (8), pp. 559-572.

Leslie, D., Mazumder, A., Peppin, A., Wolters, M.K. and Hagerty, A. (2021) *Does "AI" stand for augmenting inequality in the era of covid-19 healthcare?*, in "British Medical Journal", 372, pp. 1-5.

Maturo, A. (2015) *Doing Things with Numbers: The Quantified Self and the Gamification of Health,* in "Eä – Journal of Medical Humanities & Social Studies of Science and Technology", 7 (1), pp. 87-105.

Maturo, A. and Moretti, V. (2018) *Digital Health and the Gamification of Life: How Apps Can Promote a Positive Medicalization*, London/New York, Emerald.

McFall, L. (2019) *Personalizing solidarity? The role of self-tracking in health insurance pricing*, in "Economy and Society", 48 (1), pp. 52-76.

McFall, L., Meyers, G. and Hoyweghen, I.V. (2020) *Editorial: The personalisation of insurance: Data, behaviour and innovation*, in "Big Data & Society", 7 (2), pp. 1-11.

Moretti, V. and Maturo, A. (2021) "*Unhome" Sweet Home: The Construction of New Normalities in Italy during Covid-19*, in D. Lupton and K. Willis (eds.), *The COVID-19 Crisis: Social Perspectives*, New York, Routledge, pp. 90-102.

Neresini, F. (2015) *Quando i numeri diventano grandi: che cosa possiamo imparare dalla scienza,* in "Rassegna Italiana di Sociologia", 56 (3-4), pp. 405-431.

O'Malley, P. (1996) *Risk and responsibility*, in A. Barry, T. Osborne and N. Rose (eds.), *Foucault and Political Reason,* Chicago, University of Chicago Press, pp. 189-207.

Rowe, R. (2021) *Social determinants of health in the Big Data mode of population health risk calculation*, in "Big Data & Society", 8 (2), pp. 1-12.

Schull, N.D. (2016) *Data for life: Wearable technology and the design of self-care*, in "BioSocieties", 11 (3), pp. 317-333.

* * *

A Data Journalism Perspective on Data Circulation

Elisabetta Tola, Formicablu

Introduction

No longer in its infancy, but not a grown-up either. That is the state of data journalism, as far as Italian media are concerned. With very few exceptions, the practice of using data to provide sound, accurate, transparent information is still very rarely explored at its full potential in our country. The reasons are many, and the recent experience of the COVID-19 pandemic has made them clearer. The list is long – lack of the appropriate mindset on the side of most journalists and even more of their editors and publishers, lack of resources and skills, lack of time and reliable sources. However, most of all, lack of data. Despite the data deluge we seem to have experienced, with maps and graphs popping up everywhere, the reality is that meaningful and valuable data are still very scarce. The negative impacts and implications of such a greedy approach remain to be assessed. But there are also lessons to be learned and used to improve the information and the future journalistic work.

Data and Health: How We Learnt to Tell Stories in a Different Way

The first data-driven investigations appeared in UK and US newspapers around 2009-2010. The Guardian Datablog, founded by journalist Simon Rogers, currently data editor at Google, was one of the earliest efforts to introduce data in daily journalistic practice routinely. *Anyone can do it* – argued Rogers in a popular piece he wrote at the time: *Data journalism is the new punk*[4]. The combination of data available in easy-to-use formats released by many public institutions with the ability to use a datasheet to compile basic statistical operations and a few tools to create graphs and maps were deemed by Rogers as the basic bricks that could build a new

approach to journalism. Nevertheless, even at that initial step, it was already clear to all, and well highlighted in Rogers's article, that:

> Maybe everyone can do it, but not everyone can do it well. Like so many other things, done well is a mix of art and science (see footnote n.4).

It was almost immediately apparent that data journalism could be particularly useful when applied to issues that, while being of high public interest, are particularly difficult to interpret and understand, such as health or environmental ones. "Not all data journalism has to bring down the government – it's often enough for it to shine a light in corners that are less understood, to help us see the world a little clearer." That is Rogers, again (see footnote n.4).

One of the first iconic data investigation in the health domain was the initiative *Dollars for Docs*, a long series of online articles and a rich database built in 2015 and maintained until 2019 by the US online magazine "ProPublica"[5]. According to the methodological scaffold in which the articles are framed, the story started in 2009 when seven drug companies were required by a court to release details of their payments to doctors and teaching hospitals in the US. The story grew big in time and lasted over 10 years until, in 2019, the magazine stopped updating the database. At that point, the database included information on payments made by over 2000 companies to more than 1 million doctors and 1200 teaching hospitals for a total of over 12 billion dollars of payments. The payments included diverse categories, such as promotional speaking, consulting, meals, travels and royalties. A tool was developed that allowed any reader to search for their doctor and check whether a company had somehow paid him or her, when, how often, for what and how much. *This may put your next prescription in a different light,* wrote the editor Stephen Engelberg in one of the first commentaries accompanying the launch of the *Dollars for Docs* investigation[6]. Furthermore, the investigation and the data enacted a collaborative approach where ProPublica started co-producing articles and in-depth analyses with other media, local and national, to make sure that the database could be exploited to its total capacity and highlight many stories of local interest to citizens living in different cities and states. In many cases, it spurred actual investigations on malpractice and wrongdoing, which would have been very difficult to undertake without those data.

Shortly after that, Wired Italy launched the *Dove ti curi* [Where do you go for your health care, n.d.r.] investigation, based on the first release of data by the Agenas, the Italian National Agency for Regional Health Systems. Agenas produces a periodic report on health care quality in all 1200 Italian public hospitals. These data were made available for the first time to journalists and professionals in the health sector in 2012. The following year, Wired Italy decided to publish the entire dataset in a searchable format and many articles explaining the meaning of those data. It was an

absolute novelty in the landscape of health journalism in our country. The investigation is, unfortunately, no longer available online. However, it remains a landmark in the way health data could and should be used in enabling people to make informed decisions on where to go to be assisted, which hospitals are better for one or the other type of treatment, and which are absolutely to avoid because their rate of mortality is a dangerous outlier. Thus, those data would be precious also to researchers and professionals in the health sector for assessing and comparing the performances of the different health centres. They can show how organisational choices can have a positive impact in terms of outcomes, and they can help re-modulate or eliminate the situations where the outcomes are adverse, reducing suffering and saving lives in the end.

Interlude: How Data Enter the Journalistic Practice and Become Stories

Far from being magic, data journalism is done following a very straightforward methodology:

> Data journalism begins in one of two ways: either you have a question that needs data or a dataset that needs questioning. Whichever it is, the compilation of data is what defines it as an act of data journalism. (Bradshaw 2011).

The first step is finding and compiling the data. One can do this by finding a ready-to-use spreadsheet online, but also by using advanced scraping techniques to get data from online pages and databases, by extracting numbers from .pdf or other formats into a table, or by pulling the data using an API or finally collecting them manually, either by observations, surveys, questions, investigations. Once the database is available, it needs to be cleaned, filtered, combined, and analysed. Even with a ready-to-go table, such as the ones that can be downloaded from any open data repositories, as the data warehouses of the National Institutes of Statistics or the international organizations (such as FAOSTAT, World Bank Open Data or OECD data), the numbers are not ready to be used to write a story. This is even more true if the data are raw, and the database has been built from scratch. In this latter case, there is the need to run some statistical test to make sure that the data are significant with respect to the initial hypothesis. When the dataset has been validated, there are more operations to go through. We need to be sure that there are no mistakes, duplications, misspellings, missing information and so on, and we can achieve this step by means of running tools that highlight those errors and allow corrections in order not to misinterpret the final result.

At this point, a journalist will look at a data set in a different way than a researcher: the questions that come to mind when using data to craft an investigation are motivated by the interest in finding an angle, an explanation, an interpretation for a story that will have an impact on people's view of a certain subject. Journalists look for outliers to see if there is a moment in time, or a situation compared to others, that could be explained by external factors. Or they look for trends comparing situations that might give rise to a wider view on a certain phenomenon. Therefore, it is particularly important to be very accurate in the cleaning step since outliers are often simply the results of errors, generated either during the data collection or the data entry into the database. Comparisons in time or across different geographical situations, for instance, can only be made if the data are consistent, if the same methodology has been used to collect them in the dataset.

Furthermore, data need to be put in context, with the appropriate metadata explaining the methodology and the significance behind the data collection and organisation. In the journalism domain it is possible to use data collected in different ways, if those are the only one available, but the extent and meaning of those discrepancies have to be made clear to the readers in order to be truly informative and not misleading.

For a journalist an interesting set of data can also be a dataset that is missing a key information: at that point, the question is why that piece of information is not available. Sometimes the story can be, as a matter of fact, in the missing data, since that absence is telling something about inefficiency, malpractice, opacity and much more. Finally, in order to tell a full story, the concerned database might need to be combined with other data, such as demographic ones, historical, environmental. By addressing specific questions through the database, the journalist might see if there are interesting correlations, i.e., identifying factors that might influence or affect a certain trend. For instance, *The hunger profiteers*, a recent investigation published by Lighthouse Reports, a European collaborative investigative journalistic effort, has focused on the dramatic increase in food prices in recent months[7]. The current narrative, both by media and by many key actors and public institutions is that this increase has a lot to do with the Russian invasion of Ukraine and its impact on grain production and trade. And yet, looking more closely to the food price index estimated by Food and Agriculture Organization of the United Nations and comparing that with the global production of cereals in the last few years and the global demand, it becomes obvious that the price skyrocketed well before the Russian invasion and it does not seem to be linked to production nor demand but rather to other external factors. Investigating further and getting hold of the documents published on the main cereal trade exchange markets, those in Paris and in Chicago, the reporters exposed the role of investment funds and of speculative maneuvers on the price of cereals. These speculation have, as a primary effect, that of generating food insecurity for

millions of people. Of course, more data are needed to consolidate this interpretation, but this investigation shows exactly what the power and role of data journalism is, that of connecting and exposing data and facts of public interest. These stories are then often better explained using charts and visualisations, but that is not always the case, and sometimes charts can also lie (see Cairo 2019). In conclusion, doing a story with data requires profound respect for the data and the way they are collected and analysed. This also may require an effort in spending some time asking the appropriate questions to the database once it has been created or obtained. On the contrary, inaccurately using the data can lead to a wrong story, or no story at all.

Covering an Emergency in a Data Void

Fast forward ten years, there has been a steady increase in the number of civic activists and data journalists that bring data into the information flow. In some cases, by collaborating with a local or national media. In others by doing their work independently online, on different platforms, and organised more or less informally. There are collaborative efforts, small communities of data journalists helping each other, training courses. Many people have learned how to manage a datasheet, perform the basic checks and operations, and convert the data into meaningful graphs, charts, and maps. In these over 10 years, we have gone from simple maps and graphs showing the data in an interactive way so that the readers could select the information they most needed or wanted to see, to very elaborate data visualizations that have become, in the worst cases, more focused on the aesthetic and decorative aspects than on the informative ones. Recent investigations see a return of simple graphs that prove to be easier to access and interpret. Data journalism units and teams have been organised in small and big media outlets in many countries, and this practice has now been integrated in the journalistic practice so that maps, charts, dashboards are produced on quite a regular basis. In Europe many collaborative networks have been working on data, such as the "European Data Journalism Network" that includes news outlets from many European countries, such as "Investigate Europe": a collective effort publishing investigations, often data-driven, in different languages. There have been massive global data investigations such as the well-know "Panama papers" published by the "International Consortium of Investigative Journalists", that have been awarded the Pulitzer Prize and have seen more than 400 journalists and investigators from over 80 countries cooperating together. But also, more regional efforts, such as for example that of "Grand theft Europe", coordinated by the "German outlet Correctiv", where 63 journalists from 30 countries worked to expose the largest tax fraud in Europe perpetrated by criminal organisations, or "Don't miss the train", coordinated by "Journalism ++" and the "EDJNet".

Contrary to most other countries, with few exceptions Italian media still do not have a data team embedded in the newsroom. While the main interest of data journalism is finding new stories, new angles, and new or better explanations in the data by analysing them thoroughly, most Italian media still use the data without questioning their quality. Data are turned into something decorative, a chart here and a map there to catch the eye. Even worst, data might be cherrypicked to support a theory, a thesis, or an argument. Of course, there are numerous exceptions and truly thorough investigations done mainly by freelancers, and published on Italian media outlets. "Infodata", the data journalism section of Il Sole 24 Ore, or Wired Italy, are among the few ones to do data-driven journalism on a regular basis.

At the beginning of the COVID-19 pandemic, things changed dramatically. Before then, any investigation by media usually used the data from months or even years before. Most public institutions were (and are) still very far from releasing the raw data as a flowing stream as soon as they collect them. Hiding behind the idea that non-experts cannot access the data because they would not know how to read them, institutions demand time to clean, polish and harmonise the data before making them available to the public. At the end of February 2020, after news broke about the first Italian COVID-19 case, it soon became evident to many people from different professional environments that fresh data were badly needed. We quickly moved into one restriction after another, without having a deep understanding of what was happening and without the numbers to support many decisions. In this respect, the only guiding North Star was the daily bulletin of the Italian Civil Protection Department, released every evening at 6 p.m., listing the number of total cases by province or region. We knew nothing about the testing scale, the availability of tests, the registration of new cases compared to the previous days, as well as the testing capability. We knew nothing of the hospital capacity, how many beds could be given to the patients suffering COVID-19. For days, and then weeks, the only thing that majority of media did was publish the bulletin as it was, a .pdf table, with no further information.

Journalists Filling the Void: Bottom-up Data Related Practice

The first ones trying to have more data were the journalists working at local media outlets in the most hit cities, Bergamo, Brescia, Varese, in the region of Lombardy. Tomaso Bassani, deputy editor-in-chief of "Varese News"; Isaia Invernizzi, at that time reporter at "Eco di Bergamo"; Cristina Da Rold, freelance data journalist at "InfoData – Il Sole 24 Ore" and her colleague Riccardo Saporiti, who is also writing for Wired Italy; the team at "Il Giornale di Brescia". These journalists, to mention a few of the most active ones, immediately started looking at those data posing the same questions discussed above: *What does this data mean?*; *How were they*

collected?; *What is left out of the databases that are being shaped to monitor the COVID-19 pandemic?*. More and more journalists and activists in the fields of open data and transparency have started asking public institutions for better data and detailed information, so as to be able to compare data coming from the different cities and regions. They also launched collective efforts to get the data from the local health agencies. Around them and together with them, different grassroots associations became vocal actors in reclaiming more data: the association "OnData", advocating for open data for over a decade; the independent "Gimbe Foundation", working on evidence-based medicine; the collective "DataNinja" group on Facebook, where many data journalists discuss daily the problems encountered in working with data. For all of them, the main point was the impossibility to inform their audiences and communities helpfully because the numbers offered in the daily bulletin were meaningless. Tomaso Bassani (deputy editor-in-chief of the newspaper Varese News), for instance, started building a longer-term series, creating his own database, collecting the data daily and showing the trend in the mid and long term instead of offering just the day-to-day numbers. On March 4th 2020, OnData opened a repository on GitHub to collect, in a machine-readable format, the data published by the Civil Protection in .pdf, so journalists and activists could at least build their tables and do some analysis. At the same time, journalists campaigned to ask the Ministry of Health and the Civil Protection for the release of all the data in an open format, machine-readable, with less aggregated data needed to perform local analysis. Finally, the Civil Protection adopted the same attitude and created a GitHub repository to publish their daily data in an open format and with an open license. Since then, thousands of people have used it, showing its potential for scrutinizing the COVID-19 pandemic.

In those early weeks, journalists were among the first ones to complain about the lack of data. Those working in smaller towns were asked many questions directly from their readers regarding the actual scale of the emergency. They put up an enormous endeavor to find out more data using the old method by calling public hospitals, the regional health agencies, and the local health agencies, thus shaping databases from scratch. Their stories highlight the complete absence of an institutional culture of transparency for what concern the release and use of data. Some local health agencies or governments understood the importance of disclosing the extent of the emergency, even if only to gain support from the population. But many preferred to remain silent, hiding behind the fact that they were dealing with an emergency and did not have the time and resources to work on the data. One of the major problems was that the Italian Regions are formally in charge of locally addressing and regulating relevant issue health system. And each region works differently. We have nineteen administrative regions and two autonomous provinces, meaning 19 regional health systems and 2 provincial ones. And not only the health systems are very diverse, but

the way data are collected differs, and the results are not always comparable. Therefore, aggregating them in one table is meaningless, and it can hardly lead to any conclusion.

Besides journalists, researchers became highly interested in the data too. Given the scale of the COVID-19 pandemic, this was a unique opportunity to make in-depth data analyses at many different levels: performances of the health system; comparisons, trends, correlations to understand if the spread of the virus could also be worsened by other factors, such as environmental ones, among many others. Even local administrators and authorities can benefit greatly from the access to the data with the aim to assess the evolving situation and the measures to be enforced. An example can elucidate this point. The Italian schools remained closed for months, even in areas where the cases remained very low during the first and the second wave. The absence of data regarding the impact of the COVID-19 pandemic within the teaching system is probably still one of the less acceptable outcomes of the entire story. Studies published after the first lockdown, such as the "OpenPolis"[8] series on educational poverty or the report by "Save the children"[9] on the same topic, showed a consistent increase in learning and educational inequalities worsened by the complete unpreparedness of the Italian school system to use digital schooling appropriately and equitably. Beyond any wishful thinking, there is no doubt that our defeat to protect the most vulnerable groups of the youngest generations and to offer them a viable opportunity to attending school could have been mitigated if we had known better how the virus was spreading in schools. Reality is that even the institutions that should be on the forefront of data collection seem to lack either suitable methodologies and standards or a set of procedures in place to make those data promptly available to researchers as well as to the public, as it should be granted within an open democracy. Particularly, in the case of health data and of school-related data, each Italian region is responsible for the monitoring of the situation, for the data collection and for the communication of those data both to the central state authorities, such as the Ministry of Health or that of Education, and to the citizens through the websites.

This fragmentation has been used to justify, for instance, the inexistence of a complete database of all Italian public schools on the Ministry of education website. Only in 2019, after more than seven years of public requests, campaigns and investigations, those data have been finally made available. Therefore, what happened during the COVID-19 pandemic is not a surprise, but it is still unacceptable.

The Civil Protection daily bulletin failed to provide an accurate picture about the real death toll of the pandemic. Indeed, many mayors of local villages and towns, particularly in the most hit places (such as the province of Bergamo), highlights the low reliability and the scarce heuristic power of the official statistics on the progress of the pandemic.

With no fresh mortality data available from ISTAT – traditionally released only every three months – it was difficult to make comparisons with average mortality for specific geographic areas. That is why the local newspaper "L'Eco di Bergamo" launched its own investigation. Supported by a data science startup, the journalists designed a survey to collect the data from the local administrations, one by one. The results were distressing. In an online newspaper article published in L'Eco di Bergamo, the journalist Isaia Invernizzi argues:

> What the official figures don't say. They don't say that in March 2020 more than 5.400 people have died in Bergamo province, 4.500 of which due to Coronavirus. Six times more than the previous year. Of only 2.060 of them, the «official» certified deaths caused by COVID-19 in the local hospitals (data as at yesterday), we know everything: age, gender, pre-existing conditions. We do not know anything about the other 2.500. Many of them are old people, who died at home or in assisted residential homes. In spite of the unmistakable symptoms, as recorded by physicians and relatives, they were never tested for the disease. On their death certificate you can just read: interstitial pneumonia[10].

In this case, the data making the difference were the missing ones. Behind those data, Isaia Invernizzi and his colleagues found the most crucial story and managed to give back to those neglected dead people the right to be remembered.

Another missing piece of information, not evident at first, was the impact of COVID-19 on ordinary health care treatments. The journalist Riccardo Saporiti – supported by a scholarship granted by SISSA – Scuola Internazionale Superiore di Studi Avanzati and funded by the writer Paolo Giordano – worked for the whole year on an investigation called *Pazienti dimenticati* [Forgotten patients] (Saporiti 2021). His effort focused on the screenings, diagnostical exams, oncological treatments or other surgeries that have been cancelled or postponed due to the reorganisation of the hospitals during the COVID-19 pandemic. These postponements resulted from a political decision endorsed by the Ministry of Health in March 2020. According to Saporiti (2021):

> A decision, the judgment on which it is left to the reader, which has affected the national territory in a homogeneous way, in a context in which the pandemic has hit the country in a way that is anything but homogeneous.

Since data about the cancellation and/or rescheduling of ordinary care treatment were not publicly accessible, Saporiti had to send 200 Freedom of Information requests to Local Health Agencies and hospitals for obtaining the concerned data: 57 ignored the request; 21 rejected it; 122 sent the requested data, although not always in complete form. The request was

aimed to access to data relating to surgical interventions, outpatient visits and examinations and oncological services performed and postponed between March 1st and April 30th 2020.

In the words of Saporiti (2021), the numbers offer only:

> a photograph, albeit partial, as detailed as possible of the impact that the pandemic containment policies have had on patients not affected by Sars-CoV-2. The effects of these postponements are still all to be assessed.

The use of Freedom of Information requests and a thorough collection of available data published in scientific journals and on a range of different institutional websites were also the tools used by Davide Mancino for his one year-long investigation, called *The Big Wave* (available both in Italian and English[11]), on the health, economic and social impacts of the COVID-19 pandemic in Italy.

Elusive Data and the Campaign to Free Them

Difficulties relating to the collection and rapid release of data characterized both the first (25th February 2020 – 31st May 2020) and second (10th October 2020 – 31st December 2020) pandemic waves in Italy. During the second wave, a new system for managing the pandemic was put in place, where the different regions were assigned a colour, from yellow to red, depending on 21 parameters, the most important one being the occupation rate of hospital intensive care units. However, the whole set of 21 parameters remained very complicated to be understood by concerned people. There were weekly reports published by The Italian National Institute of Health (ISS) and the Ministry of Health, but

> if the idea was that of sharing the choices with the citizens, the result is a very complex document, comprehensible only to professionals, between numbers that do not find any explanation and algorithms that refer to previous publications (Da Rold 2020).

Citizens can no longer be expected to trust the government and institutions simply without understanding the evaluations that assign a color to each region corresponding to different levels of restrictions. According to Da Rold (2020):

> *Trust us* is no longer sufficient. Many months of sacrifices have passed, and now citizens and all those who work with data demand to know the data behind the decisions and the risk assessment.

November 2020 marked two critical steps on the data front. The first was an agreement signed between the "Accademia dei Lincei" [The Lincean Academy], whose President at the time was the Nobel Prize physicist Giorgio Parisi, with the National Institute of Nuclear Physics (INFN) and the ISS. The agreement implied that all data produced by the ISS would be given to the Accademia dei Lincei to be made available through a new platform. However, it was not clear which data were part of the bundle, and it took many other months just to see the data. Researchers from other institutions complained and criticised the decision, claiming that it would have been more fruitful to make the data available to the entire scientific community to multiply the research potential. There was also a growing interest in these data outside the scientific community. This movement finally brought to the launch of the campaign *Dati bene comune* [Data as Common Good], promoted by "ActionAid", "Ondata" and "Transparency International". The campaign was meant to foster a

> culture of open data among the Italian civil society and the public institutions and to ask the Italian government to publish open data on the management of the COVID-19 pandemic[12].

By the end of 2021, the campaign had collected more than 50.000 signatures and the support of over 275 organisations. Some results were obtained, i.e., the change of license on publications and data available on the websites of the ISS. Nevertheless, so much more needs to be done. The campaign, not limited to COVID-19, is currently asking for the data in compliance with the Recovery plan for Europe – NextGenerationEU and the application of the due economic measures.

Lessons Learned, Looking Ahead

In conclusion, the lessons to be learned are different and quite significant. In the absence of preparedness, data are the most vital tool to support decisions and try to assess risk. Data are the key ingredient to building a common ground of trust and dialogue among institutions operating at different levels and between institutions and citizens. They are the only way to promote accountability on all parts: to see if the political decisions are followed through and if the results are coherent with the premises. Furthermore, they serve the purpose of monitoring in real-time and adjusting when things go wrong. Dealing with a pandemic, as much as with an economic or an environmental crisis, requires the capacity to embrace uncertainty and complexity simultaneously. It requires a sincere and transparent attitude. The sense of frustration experienced by different stakeholders and concerned groups of people hit by the consequences of the pandemic could have found at least partial solace in the knowledge that the decisions

were taken upon precise solid data and not based on ineptitude or political calculation.

Finally, to answer those who think that data should only be handled by a restricted circle of experts and not by lay people, many experts work in different capacities and, when there is transparency, researchers, activists, journalists can indeed, independently or collaboratively, confirm or dispute calculations, interpretations and conclusions with better outcomes for the entire community. When transparent and available to all, data cannot easily be manipulated or misinterpreted or used to support wrong theories and false conclusions. A democratic and responsible society is a society where all have access to information the same way to make proper decisions, be responsible citizens, and be an active part of the joint effort to solve collective problems. Data *per se* are only one of the components of information, but in a society that is so intensely data-driven data become a very critical ingredient of a complete, transparent and honest information. Without data there is no transparency, and without transparency democracy is at risk.

References

Bradshaw, P. (2011) *The inverted pyramid of data journalism*, in https://onlinejournalismblog.com/2011/07/07/the-inverted-pyramid-of-data-journalism/ (retrieved April 14, 2022).

Cairo, A. (2019) *How Charts Lie: Getting Smarter About Visual Information*, New York, Norton & Co Inc.

Correctiv (2019) *Grand theft Europe*, in https://correctiv.org/en/top-stories/2019/05/07/grand-theft-europe/ (retrieved May 20, 2022).

Da Rold, C. (2020) *I dati della discordia sulla pandemia in Italia*, in "Le Science", https://www.lescienze.it/news/2020/11/19/news/covid-19_dati_open_pandemia_casi_morti_terapia_intensiva_criteri_chiusure_rt_calcolo_rischio_previsione-4838180/ (retrieved June 3, 2022).

EDJNET, *Don't miss the train*, in https://www.europeandatajournalism.eu/eng/Investigations/Don-t-Miss-the-Train (retrieved May 20, 2022).

Engelberg, S. (2010) *Editor's note: Dollars for Docs*, in https://www.propublica.org/article/editors-note-dollars-for-docs (retrieved May 20, 2022).

Saporiti, R. (2021) *Pazienti dimenticati*, in https://forgottenpatients.com/ (retrieved May 21, 2022).

Notes

[1] *How companies learn your secrets*, New York Times, Feb. 16, 2012, available at: https://www.nytimes.com/2012/02/19/magazine/shopping-habits.html (retrieved June 3, 2022).

[2] *How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did*, Forbes, Feb. 16, 2012, available at: https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/ (retrieved June 3, 2022).

[3] *The world after the Coronavirus*, Financial Times, April 19, 2020, available at: https://www.ft.com/content/19d90308-6858-11ea-a3c9-1fe6fedcca75 (retrieved June 3, 2022).

[4] *Anyone can do it. Data journalism is the new punk*, in https://www.theguardian.com/news/datablog/2012/may/24/data-journalism-punk (retrieved April 14, 2022).

[5] See: https://www.propublica.org/series/dollars-for-docs (retrieved June 3, 2022)

[6] See: https://www.propublica.org/article/editors-note-dollars-for-docs (retrieved June 3, 2022).

[7] See: https://www.lighthousereports.nl/investigation/the-hunger-profiteers/ (retrieved June 3, 2022).

[8] See: https://www.savethechildren.it/cosa-facciamo/pubblicazioni/impatto-del-coronavirus-sulla-poverta-educativa (retrieved May 20, 2022).

[9] See: https://www.openpolis.it/poverta-educativa/ (retrieved May 20, 2022).

[10] See: *Coronavirus, the real death toll: 4500 victims in one month in the province of Bergamo*, in https://www.ecodibergamo.it/stories/bergamo-citta/coronavirus-the-real-death-tool-4500-victims-in-one-month-in-the-province-of_1347414_11/ (retrieved April 14, 2022).

[11] See: https://www.grandeonda.it/en/ (retrieved May 20, 2022).

[12] See: https://www.datibenecomune.it/advocacy/.