The next Gold Rush: Medical and health data

Anurag Mehra

The messy, digital data-rich universe that is emerging rapidly is being nurtured and bolstered by powerful tech companies. Whatever the potential benefits for human welfare, the development is posing serious dilemmas that need urgent resolution.

Health data, especially patient medical records, in the form of Electronic Health Records (EHRs) are the elixir that Artificial Intelligence/Machine Learning (AI/ML) systems have been waiting for. The idea is to use this humongous amount of data to create and train algorithms that can find new patterns which can lead to improved diagnosis, prognosis and outcomes. Detailed records contain doctor's notes which can be invaluable in developing these technologies. The main obstacles to such a grand venture are fragmentation of standards used to create such records, the high costs involved, and prickly questions of data ownership and patient consent.

The Covid-19 pandemic emergency seems to have opened up data records for processing. The *Economist* (*https://www.economist.com/science-and-technology/2020/05/14/the-pandemic-has-spawned-a-new-way-to-study-medical-records*) reports a study in the United Kingdom (U.K.), based on the data available with the National Health Service (NHS), on the patterns that characterize the profiles of people who died because of the disease. The investigation examined health profiles of around 5700 people who lost their lives to the covid disease, out of a reference base of nearly 17 million people in NHS records. The results provide the first, large-scale validation of the observations that have often been discussed in the public sphere, that: males die more than females; the old are at greater risk than the young; persons with other pre-existing diseases are more vulnerable; and, ethnic minorities (blacks, Asians) have more fatalities than whites. However, what deserves comment are three things.

First, is the method by which the patient data in the NHS database was accessed. The NHS database is amongst the largest in the world, with all data being reported at the "family physician" (General Practitioner - GP) level. Each record gives a detailed (longitudinal) picture across time, starting with the first report of a sickness to a GP and then along the illness trajectory deeper into specialist domains. This is recorded for every episode of falling ill. The data is stored in a standardized form across the whole country, and is managed typically by a records vendor. The data analysis was performed using software tools that could access and condition the data (e.g. make the data pseudo-anonymous) within the database, and without copying the records anywhere else. The OpenSAFELY website states, "We don't transport large volumes of potentially disclosive pseudonymised patient data outside of the secure environments

Anurag Mehra teaches engineering and policy at IIT Bombay. His policy focus is the interface between technology, culture and politics.

managed by the electronic health record software company; instead, trusted analysts can run large scale computation across live pseudonymised patient records inside the data centre of the electronic health records software company".

This connects us to the second feature of this study, the issue of establishing trust that allowed the project to access the data. The analytics platform OpenSAFELY has "created trust" in many ways. The platform is developed as a collaborative effort primarily between entities that command a great deal of public respect: DataLab at the University of Oxford (specialty in medical data projects), the EHR group at London School of Hygiene and Tropical Medicine (EHR research) and the major records management vendor TPP (records-related software expertise). The project is led by charismatic NHS doctors possibly with political heft.

Lastly, the team has addressed the problem of making the access process secure and transparent by keeping logs of all the activity undertaken during the process of analysis at each step. This is very welcome oversight of the transactions carried out. Further, the analytic tools are open for scientific and security review, and for reuse as open source software.

Even though the "quick" permissions to access NHS data have been driven by the Covid-10 pandemic, this is a pioneering effort, at scale, and presents a possible model for analysing confidential health records in the best possible manner. It should remind us of the benefits of having a robust and sensible public health system, quite a contrast to the neo-liberal dream of having an American style, private insurance-based health care system.

Yet, a project like this, may also set a precedent, so that someone could demand access to health records but not necessarily with the same level of safeguards or credibility. The NHS itself has been found wanting on many fronts in this regard. There were many misgivings and reservations when NHS data was first pushed into centralized databases, in 2014. In fact, an op-ed piece in *Nature,* found assurances about data safety and privacy too glib and overplayed, and pointed out that patient opt-out was an important issue; this "care.data" project was ultimately abandoned. However, NHS data, because of its scale and standardization, is considered a gold mine to which many seek access. So attempts continue to seek as much of it, by multinational pharma companies as well as Big Tech from Silicon Valley.

For a few years now, the Department of Health and Social Care has been selling "anonymized" NHS data to international drug companies, such as, Merck, Bristol-Myers Squibb and Eli Lilly, for neat sums of money, for research. These sales have happened amidst a huge furore over how anonymous really is the data, that has been sold. It has been argued, and rightly so, that much of this data can be easily linked back to individual patients' records. In fact, it has been suggested that for some patients of special interest the buyer companies have already de-anonymized the data they purchased! Of course, it is tremendously naive to believe that just stripping off names or enrollment numbers from health records will make the data anonymous. This is simply because the longitudinal nature of a record - containing a full history of illnesses and interventions - make them as unique as fingerprints. The interest of the American government in ensuring that NHS data is accessible to US drug majors is evident from its stance about post-Brexit agreements and trade deals.

The other bounty hunters in this game are Big Tech companies such as Amazon, Google, Facebook et al. Each of them has a different kind of interest in medical data. Amazon does not have access to patients' records (yet!) but has signed a deal with the NHS, which allows it to access healthcare information "collected by the NHS and displayed on its website". The goal is to enable Amazon's digital (audio) assistant, Alexa, to be able to answer basic health-related questions. This service, the UK government thinks, will take the load off the GPs network. Though the deal was signed in 2018, no evidence is available to show whether such a load reduction has happened or not. More importantly, we do not know if Alexa is able to hold complex medical conversations with customers. Amazon is, of course, free to advertise this as a selling point for Alexa-equipped devices. The problems that this situation creates are that it allows Amazon to develop a kind of medical profile not just for an individual but as more people pose their queries the data can be aggregated and analysed for patterns across groups.

Google has been trying to build e-health repositories for very long. These will be added to all the other data they have about us and ensure that our google profiles also have sections devoted to health-related information. But more than that, health data will allow Google to train its AI/ML systems and to mine it for disease patterns, vulnerabilities, cures, efficacies and more. Google's acquisition of FitBit, the fitness tracker company that makes wearables, is also an effort to be comprehensive in its health focused initiatives. FitBit has fitness data of 28 million users and this merger will put Google in a near-monopolistic advantage. The acquisition has run into regulatory hurdles in many countries, and at this point of time Google has promised that health data will not be used for advertising services. Google already has a history of accessing NHS data.

It is a matter of time before other tech giants too make a grab for NHS data. Facebook has been a late entrant into the health data game with the launch of its Preventive Health tool. The surreptitious 'influence' of big tech should not be underestimated. As recently as just about a year ago it was reported that ad-trackers on NHS websites were sending data about queries put in by users, to Google and Facebook!

It may not be surprising to see Big Tech demanding a "within-the-warehouse" access to NHS data on the lines of the OpenSAFELY initiative. Given that, with the exception of China, Big Tech dominates and monopolizes AI/ML initiatives, they will be able to offer cutting edge tools to do even more of this type of analysis. The point to ponder about will be: what will they take out of the data warehouse, as their pound of flesh?

So here we are in a messy, digital, data-rich universe, created and nurtured by powerful technology companies. In the context of healthcare, there are two clear dilemmas: first, how do we obtain the massive benefits of analysing large scale health data while respecting patients' rights to privacy and data security; second, how do we ensure that these benefits accrue to universal healthcare programmes (especially, public health systems) even as we institute a 'fair' regulation of profits and intellectual property claims made by powerful private corporations.