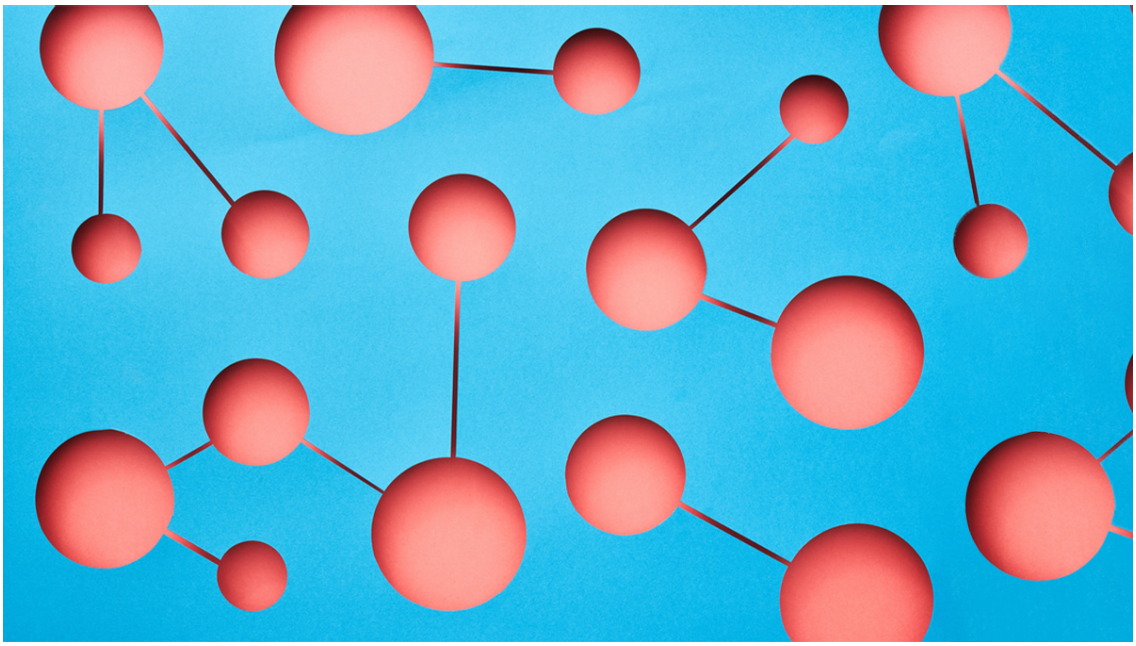
Harvard
Business
Review

Analytics And Data Science

Why AI Failed to Live Up to Its Potential During the Pandemic

by Bhaskar Chakravorti

March 17, 2022



Martí Sans/Stocksy

Summary. The pandemic could have been the moment when AI made good on its promising potential. There was an unprecedented convergence of the need for fast, evidence-based decisions and large-scale problem-solving with datasets spilling out of every country in... **more**

The Covid-19 pandemic was the perfect moment for AI to, literally, save the world. There was an unprecedented convergence of the need for fast, evidence-based decisions and large-scale problem solving with datasets spilling out of every

country in the world. For health care systems facing a brand new, rapidly spreading disease, AI was — in theory — the ideal tool. AI could be deployed to make predictions, enhance efficiencies, and free up staff through automation; it could help rapidly process vast amounts of information and make lifesaving decisions.

Or, that was the idea at least. But what actually happened is that AI mostly failed.

There were scattered successes, no doubt. Adoption of automation picked up in retail warehouses and airports; chatbots took over customer service as workers were in lockdown; AI-aided decisions helped narrow down site selections for vaccine trials or helped speed up border crossings in Greece.

In general, however, in diagnosing Covid, predicting its course through a population, and managing the care of those with symptoms, AI-based decision tools failed to deliver. Now that some of the confusion of the pandemic's early days has settled, it's time to reflect on how AI performed on its own "Covid test." While this was a missed opportunity, the experience provides clues for how AI systems must evolve to realize the elevated expectations for what was the most talked about technology of the past year.

Where AI Failed

At the outset, things looked promising. Machines beat humans in raising the early alert about a mysterious new virus out of Wuhan, China. Boston Children's Hospital's HealthMap system, which scrapes online news and social media for early signals of diseases, along with a Canadian health news scraper, BlueDot, picked up warning signs. BlueDot's algorithm even predicted cities most at risk if infected people were to travel, all days before the WHO and weeks before the rest of the world caught up.

As the world officially went into lockdown in 2020, it was clear that AI's game-changing contribution would be in rapid

prediction — diagnosis, prognosis, and forecasting the spread of an emergent unknown disease, with no easy way to test for it in a timely way.

Numerous AI-enabled teams mobilized to seize the opportunity. At New York's Mount Sinai hospital, for example, a team designed an AI system to quickly diagnose Covid-19 using algorithms trained on lung CT scans data from China. Another group at MIT created a diagnostic using algorithms trained on coughing sounds. A third team, an NYU and Chinese collaboration, used AI tools to predict which Covid-19 patients would develop severe respiratory disease. We had heard for years about AI's transformative potential, and suddenly there was an opportunity to see it in action.

So, how did these AI-powered Covid predictors work out? Put bluntly, they landed with a thud. A systematic review in *The BMJ* of tools for diagnosis and prognosis of Covid-19 found that the predictive performance was weak in real-world clinical settings. Another study at the University of Cambridge of over 400 tools using deep-learning models for diagnosing Covid-19 applied to chest x-rays and CT scans data found them entirely unusable. A third study reported in the journal, *Nature*, considered a wide range of applications, including predictions, outbreak detection, real-time monitoring of adherence to public health recommendations, and response to treatments and found them to be of little practical use.

We can learn from these disappointments as we gear up to build back a better AI, however. There are four places where the fault lines appeared: bad datasets, automated discrimination, human failures, and a complex global context. While they relate to Covid-19 decisions, the lessons are widely applicable.

The Danger of Bad Datasets

AI decision-making tools are only as good as the data used to train the underlying algorithms. If the datasets are bad, the algorithms

make poor decisions. In the context of Covid, there are many barriers to assembling “good” datasets.

First, the breadth of Covid symptoms underscored the challenge of assembling comprehensive datasets. The data had to be pulled from multiple disparate electronic health records, which were typically locked away within different institutional systems and their corresponding siloes. Not only was each system separate, they also had different data governance standards with incompatible consent and confidentiality policies. These issues were amplified by health care systems spanning different countries, with incompatible patient privacy, data governance, and localization rules that limited the wholesale blending of such datasets.

The ultimate impact of such incomplete and poor-quality data was that it resulted in poor predictions, making the AI decision tools unreliable and untrustworthy.

A second problem arose from the way data was collected and stored in clinical settings. Aggregated case counts are easier to assemble, but they may omit key details about a patient’s history and other demographic, personal, and social attributes. Even finer details around when the patient was exposed, exhibited symptoms, and got tested and the nature of the symptoms, which variant they had been infected with, the medical interventions and their outcomes, etc., are all important for predicting how the virus might propagate. To compound the problems, some datasets were spliced together from multiple sources, introducing inconsistencies and redundancies.

Third, a comprehensive dataset with clues regarding Covid symptoms, how the disease might spread, who is more or less susceptible, and how to manage the disease ought to draw from multiple sources, given its newness. In addition to data from the formal health care settings, there are other critical information sources, datasets, and analyses relevant for predicting the

pathways of a novel and emergent disease. Such additional data may be drawn from multiple repositories, effectively tapping into the experiences of people grappling with the disease. Such repositories could include Twitter, professional message boards, analyses done by professionals and amateurs on "open-source" platforms, medical journals, blogs, and news outlets. Of course, once you account for so many disparate sources of relevant data, the process of integration, correcting for wrong or misinformation, fixing inconsistencies, and training algorithms increased the complexity of creating a full dataset.

Automated Discrimination

Even when there were data available, the predictions and decisions recommended by health care management algorithms led to potentially highly discriminatory decisions — and concerns that some patients received worse care. This is because the datasets used to train the algorithms reflected a record of historical anomalies and inequities: lower levels of access to quality healthcare; incorrect and incomplete records; and deep-seated distrust in the health care system that led some groups to avoid it.

There are broad concerns about the negative impacts of AI bias, but during the pandemic, the consequences of such bias were severe. For example, consider a pre-Covid study in *Science* that found that Black patients were assigned the same risk level by an algorithm as white patients, even though the latter were not as sick — leading to inadequate medical care for the Black patients. Looking ahead, as Black and Hispanic Covid-19 patients suffered higher mortality rates than white patients, algorithms trained on such data could recommend that hospitals redirect their scarce resources away from Black and Hispanic patients.

The ultimate impact of such automated discrimination is even more distortionary when we consider that these disadvantaged groups have also been disproportionately affected by the most

severe cases of Covid-19 — in the U.S., Black, Hispanic, and Native Americans were about twice as likely to die from the disease as white patients.

Human Error

The quality of any AI system cannot be decoupled from people and organizations. Behaviors, from choosing which applications and datasets are used to interpreting the decisions, are shaped by incentives and organizational contexts.

The wrong incentives can be a big problem. Managers overseeing health care systems often had few incentives to share data on patients — data may have been tied to revenues, or sharing it may raise concerns over patient confidentiality. For researchers, rewards were often aligned with sharing data with some select parties but not everyone. Moreover, there were few career incentives to validating existing results, as there is greater glory in producing new findings rather than replicating or validating other studies. This means that study results may not have applied in a wide enough variety of settings, making them unreliable or unusable and causing caregivers to hesitate to use tools that had not been proven in multiple settings. It is particularly risky to experiment with human health.

Then, there's the issue of data entry errors. Much of the data accumulated on Covid-19 involved environments in which health care workers were operating under pressure and extraordinarily heavy caseloads. This may have contributed to mislabeled and incomplete datasets — with mistakes showing up even in death certificates. In many countries, health care systems were underreporting Covid-19 cases, either because they were encouraged to do so by the authorities, because of unclear guidelines, or simply because staff were overwhelmed.

Even with AI tools on hand, the humans responsible for making decisions often lacked critical interpretive capabilities — from language to context awareness or the ability to spot biases and

mistakes. There isn't, as yet, a uniformly accepted code of ethics, or a checklist, that gives caregivers a sense of when to apply AI tools versus mitigating harms by using judgment. This could lead to inconsistent use or misuse of the AI tools and eventually undermine trust in them.

Complex and Uneven Global Context

A pandemic, by definition, cuts across different political, economic, and sociocultural systems. This complicates the process of assembling a comprehensive dataset that aggregates across different countries with widely applicable lessons. The pandemic underscored the challenge of deriving universally applicable decision tools to manage human health across all health care settings regardless of geographic location. Appropriate medical interventions depend on many factors, from biology to institutional, sociopolitical, and cultural forces to the local environment. Even if many facets of human biology are common across the world, the other factors vary widely.

For one, there are differences across countries in terms of their policies regarding data governance. Many countries have data localization laws that prevent the data from being transported across borders. There is no international consensus on how health care data should be shared. While the preexisting international network for the sharing of influenza genome sequence data was extended to the sharing of sequences for Covid-19, deeper data-sharing collaborations between countries could have helped with ongoing management of the disease. The absence of broader sharing agreements and governance was a critical barrier.

Second, there were differences between developed and developing countries regarding sharing of health care data. Some researchers argue that genome sequences should be shared on open databases to allow large-scale analyses. Others worry about exploitation; they are concerned that researchers and institutions from poorer countries weren't given adequate credit and the

benefits of the data sharing would be limited to rich countries.

Third, history and the sociopolitical contexts of countries and their ethical frameworks for data sharing even within their own citizenry are different, giving rise to differences in the willingness to have personal data collected, analyzed, and shared for public use. Consider the varied experiences with AI-aided exposure identification and contact tracing apps.

South Korea presented an extreme example of intrusive data collection. The country deployed contact tracing technology together with widespread testing. Its tracking apps were paired with CCTV footage, travel and medical records, and credit card transaction information. Koreans' willingness to tolerate this level of intrusion can be traced to the country's history. The previous administration had botched its response to the 2015 MERS outbreak, when it shared no information about hospitals visited by infected citizens. This led to public support for legislation giving health authorities access to data on infected citizens and the right to issue alerts. In contrast, the German government's contact tracing app was rejected by the public once a highly critical open letter from experts raised fears of state surveillance. As a result, Germany abandoned the centralized model for a decentralized alternative. Again, history provides an explanation. Germans have lived through two notorious surveillance regimes: the Gestapo during the Nazi era and the Stasi during the Cold War. Centrally controlled state data collection was not destined to be popular.

Finally, the data on patients from one country may not be good predictors in other countries. A variety of other factors from race, demographics, socioeconomic circumstances, quality of health care, immunity levels, co-morbidities, etc., make a difference.

What to Do Now

There are several lessons to be drawn that can help improve future AI systems that must be ready for the next pandemic.

1) Find better ways to assemble comprehensive datasets and merge data from multiple sources. It would help to have health care datasets in standardized formats paired with mechanisms to create centralized repositories of data. New data-processing techniques should be considered as well. Examples include allowance for differential privacy or using synthetic data rather than real data as the technologies to facilitate such innovations improve. Moreover, the problem is not just of fragmented or incomplete data; it is also one of too much data. The transmissibility of the virus, the fact that it mutates constantly, the movement of people across borders, and the widespread use of genomic sequencing means that AI systems must deal with a deluge of data. There must be systems in place that can handle such large datasets and appropriately label and organize them.

2) There needs to be a diversity of data sources. Some lessons can be learned from the example of Nightingale Open Science, which has amassed 40 terabytes of medical imagery across a wide range of conditions and treatments along with a diversity of patient data and outcomes. These will be used to train algorithms to predict medical conditions earlier, conduct triage, and save lives in an unbiased manner. They try to work with health systems across the world, specifically including underresourced ones, to mitigate the possibilities of underrepresentation and avoid automated discrimination.

3) Incentives must be aligned to ensure greater cooperation across teams and systems. AI teams should also be provided the opportunities and incentives to collaborate with clinicians and others who are knowledgeable about the practical issues. It is also essential to plan for a diversity of stakeholder groups involved in setting ethical frameworks and checklists for practitioners using AI in mission-critical settings, along with clear processes for governance and accountability. Such groups should include engineers and technologists, experts in key functional areas, as well as ethicists who can guide the use of AI systems and their

alignment with value judgments.

Appealing to open-source communities is another way of cooperatively assembling data from multiple sources. The Open COVID-19 Data Working Group, the MIDAS Network, and other local collaborative efforts provide models that others can replicate. Enabling ways for interdisciplinary collaborations can be key to breakthroughs. For example, BioNTech, the German biotech company that pioneered the messenger RNA technology behind the Pfizer Covid-19 vaccine, has teamed up with London-based AI company InstaDeep to create an "early warning system" for spotting new coronavirus variants.

4) Write international rules for data sharing. For health data to be shared between countries, we need international conventions facilitating pooling of such critical information and agreements on data sharing, while preserving privacy and confidentiality. AI teams need to be trained to recognize differences in global health care environments, so they can place data from different parts of the world in appropriate context.

As this pandemic becomes endemic and we prepare for the next one, there are many opportunities for AI to make its mark. After Google's much-hyped Flu Trends missed the magnitude of the 2013 flu season, Covid offered a dramatic chance at redemption for AI as a predictive tool. But within the current failures lie the seeds of AI systems that can flourish in the future.

Bhaskar Chakravorti is the Dean of Global Business at The Fletcher School at Tufts University and founding Executive Director of Fletcher's Institute for Business in the Global Context. He is the author of *The Slow Pace of Fast Change*.