Machine Learning for COVID-19 needs global collaboration and data-sharing

The COVID-19 pandemic poses a historical challenge to society. The profusion of data requires machine learning to improve and accelerate COVID-19 diagnosis, prognosis and treatment. However, a global and open approach is necessary to avoid pitfalls in these applications.

Nathan Peiffer-Smadja, Redwan Maatoug, François-Xavier Lescure, Eric D'Ortenzio, Joëlle Pineau and Jean-Rémi King

On 31 December 2019, the first cases of a viral pneumonia with unknown aetiology were reported in the city of Wuhan, China. In the following weeks, the Chinese authorities and the World Health Organization (WHO) announced the discovery of a novel coronavirus and its associated disease: SARS-CoV-2 and COVID-19, respectively. On 21 April 2020, the number of cases of COVID-19 exceeded 2.4 million and the death toll exceeded 170,000 worldwide[1]. The outbreak of COVID-19 represents a major and urgent threat to global health. While the unprecedented speed of the COVID-19 spread partly finds its roots in our increasingly globalized society, the global sharing of scientific data also offers a promising tool to fight the disease. In the past four months, more than 12,400 articles have been published[2] and scientific data collected from thousands of patients have been released[3]. The majority of these studies follow the standard scientific method: that is, investigate a few hypotheses at a time on a controlled sample. While undeniably successful, this standard method suffers from two well-known challenges, both critical to our pandemic situation: (1) it requires considerable expertise and human input and (2) it only considers a handful hypotheses at a time. Machine learning (ML) has been used to meet these challenges in various pathologies[4,5], including infectious diseases[6]. Herein, we describe two areas where ML could supplement standard statistical methods in the COVID-19 pandemic, discuss the practical challenges that such a ML approach entails, and advocate for a global collaboration and data-sharing.

ML to alleviate the workload of medical experts

While standard statistical methods can provide the first results necessary in an emergency, they often require considerable human resources, which are precisely lacking in such context. Health systems found themselves quickly overwhelmed and the potential for data analysis, particularly in clinical research, was limited by the amount of work required. ML techniques can decrease the time required to produce automated analyses and allow artificial intelligence practitioners to support clinicians. For example, medical imaging studies show that chest computed tomography (CT) scans can be used to detect COVID-19 lesions[7,8]. However, such studies typically require each scan to be reviewed by a trained radiologist, who could otherwise be working on the front lines. ML may alleviate this task: recent supervised classifiers trained over a large dataset of 400,000 chest X-Rays achieved a mean area under the receiver operating characteristic curve (false positive rate versus true positive rate) of 94% for the diagnosis of 14 distinct lung pathologies[9]. Furthermore, preliminary studies based on a few hundred chest CT scans suggest that COVID-19 can be automatically diagnosed with ML[10]. However, the use of ML of medical images to diagnose or prognose COVID-19 remains currently limited to relatively small cohorts. These studies thus poorly control for the numerous confounds (for example, age, corpulence) that the algorithms may detect from chest images. A promising strategy is to pre-train ML models from larger datasets of similar images, thus learning common features to compute, which can then be used to facilitate training from COVID-19 images. This strategy has been used again and again in computer vision in recent years, to achieve impressive results in tasks with few labelled examples[11].

ML to accelerate the screening of treatments

Standard methods only consider a handful hypotheses at a time. For example, among more than 1,200 clinical trials that have been registered to identify treatments for COVID-19, the majority focus on a unique drug or a couple of drugs, hand-selected on rationales of varying relevance[12]. ML can broaden such a screening and selection process by simultaneously considering several potential antiviral agents, relying on DNA sequences and/or protein structure, including potential drug binding sites of SARS-CoV-2, to predict interactions between drugs and the virus, and thus shortlisting promising candidate treatments[13,14]. ML has been used in other infectious diseases in a similar fashion[15]: for example, a deep neural network was successfully trained to screen the activity of more than 100 million molecules on *Escherichia coli*[16]. In the same way, a large spectrum of vaccine candidates could be screened based on their potential to elicit an effective immune response, for example, by presenting the spike protein S that follows a SARS-CoV-2 infection[17]. Nonetheless, these potentially fruitful avenues should not hide the challenges of therapeutic research based on ML. First, ML cannot accelerate basic biology, and even the prediction of protein folding remains a remarkably difficult problem[18]. In the case of vaccines, there is therefore a necessary waiting period. Second, a major ethical concern is the temptation to bypass proper clinical trials: working with very small cohorts, not using adequate design, or omitting inclusion and exclusion criteria have already be reported in the recent hydroxychloroquine-based treatment research[19]. This risk could dramatically increase with ML algorithms. Indeed, algorithms such as deep neural networks are 'general approximators': they can be trained to fit any objective on a dataset by, for example, memorizing the diagnosis of every patients. ML algorithms can only be evaluated conclusively by assessing their ability to accurately predict an independent test set — an approach that necessitate

large datasets and a priori inclusion and exclusion criteria.

A major need for data sharing

While standard statistical analyses are adapted for many clinical and epidemiological challenges, ML is essential to accelerate the analysis of complex and large datasets such as large genomic or medical imaging datasets. Overall, ML is thus promised to supplement rather than supersede standard methods used for diagnoses, prognosis and treatment. However, two major challenges currently limit the potential impact of ML. First, ML algorithms are notoriously difficult to interpret. While visualization tools may highlight the combination of variables that led an algorithm to make a particular prediction, healthcare professionals must be aware that, like humans, ML can easily be affected by systematic biases (for example, scanning device, patient's age and so on). Special pedagogical efforts must thus be made in both scientific reports and in the clinics to maintain a healthy scepticism when it comes to ML findings. Second, the lack of large healthcare, clinical, imaging and genetic public repositories leads each institution to locally develop its own analytical pipeline on its own small dataset, which significantly limits the generalizability of the results. While this issue is not specific to ML, the ability of modern algorithms to encompass heterogeneous datasets should drive us to both (1) share the de-anonymized raw data used in each clinical study, and (2) favour the development of large cohorts. The International Severe Acute Respiratory and emerging Infection Consortium (ISARIC) initiative aims to provide a large and shared clinical database on COVID-19 patients[20]. Other institutions have signed data-sharing agreements to ensure that data is shared widely and rapidly[21,22], and can inform new hypotheses, but this is still done in a piecemeal fashion, making it difficult to make the most of the data generated daily during the pandemic. Not only will the quality of the standard and ML models directly depend on the size, quality and representativeness of such databases, but they will be critical to support effective interventions across different countries and types of healthcare facilities[6]. Open sharing of clinical databases requires significant care to properly manage regulatory and data privacy issues. Rapidly resolving these issues can be particularly challenging during a pandemic, when many public institutions are not operating normally. However, until we meet these challenges, ML may not keep its promises to help fight the virus.

Conclusion

The COVID-19 outbreak is not the first pandemic and is unlikely to be the last. For the first time, however, our societies have the means to provide a coordinated, evidence-based, fair and global public-health response. While the efficiency of this response may partly depend on ML, it depends even more crucially on our ability to set up global collaborations and data-sharing agreements that can accelerate the discovery and validation of promising interventions. ❐

Nathan Peiffer-Smadja[1]✉, RedwanMaatoug[2], François-XavierLescure[3], Eric D'Ortenzio[4], Joëlle Pineau[5,6] and Jean-Rémi King[7,8]

[1]IAME, Inserm, UFR de Médecine Paris 7 Denis Diderot, Paris, France. [2]Pitié-Salpêtrière Hospital, Paris, France. [3]Department of Infectious Diseases, Hôpital Bichat-Claude-Bernard, Paris, France. [4]REACTing, Inserm, Paris, France. [5]McGill University, Montreal, Canada. [6]Facebook AI Research, Montreal, Canada. [7]École Normale Supérieure, PSL University, CNRS, Paris, France. [8]Facebook AI Research, Paris, France.
✉*e-mail: nathan.peiffer-smadja@inserm.fr*

Published online: 22 May 2020
https://doi.org/10.1038/s42256-020-0181-6

References

1. Dong, E., Du, H. & Gardner, L. *Lancet Infect. Dis.* **20**, 533–534 (2020).
2. Dimensions COVID-19 publications, data sets, clinical trials. *Figshare* https://dimensions.figshare.com/articles/Dimensions_COVID-19_publications_datasets_and_clinical_trials/11961063 (2020).
3. Wu, Z. & McGoogan, J. M. *JAMA* **323**, 1239–1242 (2020).
4. Claassen, J. et al. *N. Engl. J. Med.* **380**, 2497–2505 (2019).
5. Sitt, J. D. et al. *Brain* **137**, 2258–2270 (2014).
6. Peiffer-Smadja, N. et al. *Clin. Microbiol. Infect.* https://doi.org/10.1016/j.cmi.2019.09.009 (2019).
7. Ai, T. et al. *Radiology* https://doi.org/10.1148/radiol.2020200642 (2020).
8. Chen, Z. et al. *Eur. J. Radiol.* **126**, 108972 (2020).
9. Pham, H. H., Le, T. T., Tran, D. Q., Ngo, D. T. & Nguyen, H. Q. Preprint at https://arxiv.org/abs/1911.06475 (2019).
10. Zheng, C. et al. Preprint at https://doi.org/10.1101/2020.03.12.20027185 (2020).
11. Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. Preprint at https://arxiv.org/abs/2002.05709 (2020).
12. Belhadi, D. et al. Preprint at https://doi.org/10.1101/2020.03.18.20038190 (2020).
13. Liu, X. & Wang, X.-J. *J. Genet. Genom.* **47**, 119–121 (2020).
14. Computational predictions of protein structures associated with COVID-19. *Deepmind* https://deepmind.com/research/open-source/computational-predictions-of-protein-structures-associated-with-COVID-19 (2020).
15. Peiffer-Smadja, N. et al. *Clin. Microbiol. Infect.* https://doi.org/10.1016/j.cmi.2020.02.006 (2020).
16. Stokes, J. M. et al. *Cell* **180**, 688–702e13 (2020).
17. Weiskopf, D. et al. Preprint at https://doi.org/10.1101/2020.04.11.20062349 (2020).
18. Senior, A. W. et al. *Nature* **577**, 706–710 (2020).
19. Gautret, P. et al. *Int. J. Antimicrob. Agents* https://doi.org/10.1016/j.ijantimicag.2020.105949 (2020).
20. COVID-19 Clinical Research Coalition *Lancet* **395**, 1322–1325 (2020).
21. Sharing research data and findings relevant to the novel coronavirus (COVID-19) outbreak. *Wellcome Trust* https://wellcome.ac.uk/coronavirus-covid-19/open-data (2020).
22. Open-access data and computational resources to address COVID-19. *National Institutes of Health* https://datascience.nih.gov/covid-19-open-access-resources (2020).