Artificial Intelligence, Covid-19 and Law: The Need for Evaluation in the Age of Many Models

Journal:	*BMJ*
Manuscript ID	BMJ-2020-059371
Article Type:	Analysis
BMJ Journal:	BMJ
Date Submitted by the Author:	17-Jun-2020
Complete List of Authors:	Krass, Mark; Stanford University Henderson, Peter; Stanford University, Computer Science Mello, Michelle; Stanford University Law School Studdert, David; Stanford University, Medicine Ho, Daniel; Stanford University
Keywords:	Covid-19, Law, Artificial intelligence, Machine learning, Public health

Analysis

Artificial Intelligence, Covid-19 and Law: The Need for Evaluation in the Age of Many Models

Mark Krass[1,2]
Peter Henderson[1,3]
Michelle M. Mello[1,4]
David M. Studdert[1,4]
Daniel E. Ho[1,2]

[1] Stanford Law School, Stanford University, Stanford CA, USA
[2] Department of Political Science, Stanford University School of Humanities and Sciences, Stanford CA, USA
[3] Department of Computer Science, Stanford University School of Engineering, Stanford CA, USA
[4] Stanford Health Policy and Department of Medicine, Stanford University School of Medicine, Stanford CA, USA

Correspondence to:
Daniel E. Ho
559 Nathan Abbott Way
Stanford, CA 94305
Email: dho@law.stanford.edu
Phone: +1 650-723-9560

Word count: 2000
References: 20

KEY MESSAGES

- A proliferation of models using Artificial Intelligence (AI) and Machine Learning (ML) are in use or under development to predict individuals' Covid-19 related risk.
- The use of personally identifiable information, including race, raises legal concerns over privacy and antidiscrimination, which we illustrate in the context of American law.
- The underlying legal principles ultimately boil down to an assessment of effectiveness and burdens of AI/ML tools.
- More robust evaluation of AI/ML tools will be necessary to support the adoption and legality of rapidly proliferating tools.

Contributors and sources
MK, DEH, and PH conceived of this project. All authors discussed the results and contributed to the final manuscript. DEH acts as guarantor.

Patient involvement
No patients were involved in this research.

Conflicts of Interest
We have read and understood BMJ policy on declaration of interests and have no conflicts to declare.

Licence

Artificial Intelligence, Covid-19 and Law: The Need for Evaluation in the Age of Many Models

Standfirst

Daniel E. Ho and colleagues explore the legal implications of artificial intelligence (AI) to fight Covid-19 and argue that the law militates in favor of more robust evaluation frameworks given rapidly proliferating AI models.

The potential utility of artificial intelligence (AI) and machine learning (ML) in the fight against Covid-19 has sparked a flurry of proposals, prototypes, and models. Governments have been especially interested in systems that can identify, track, and manage at-risk individuals using extensive data. Examples include China's Alipay Health Code App that color-codes citizens for quarantine based on user behavior,[1] Covid-19 vulnerability predictions for individuals,[2] voice-based detection of infection,[3] computer vision for fever detection, and facial recognition for tracking individuals and assessing compliance with face mask requirements.[4]

Are these uses of AI/ML by governments legal? A key part of the answer to that question relates to the effectiveness of the models in combating Covid-19 and the burdens their use imposes on individuals. We explore those issues by reference to two of the most salient legal concerns: (i) protecting privacy, which poses challenges because reducing the amount of personally identifiable information used in algorithms can reduce their accuracy; and (ii) avoiding discrimination, which can be difficult given the differential burden of Covid-19 across age, gender, and racial groups.[5] We argue that robust evaluation of AI/ML models is required to assess their legality in these areas. This evaluation must focus on demonstrating two things: that the models produce valid, reliable predictions and that they burden individuals' civil liberties no more than necessary. In evaluating the legality of public health officials' use of algorithms, courts will likely go beyond this, also inquiring into how the output of these tools is used to shape policies and programs. But showing that a model performs well and does not burden privacy and other interests more than it needs to are essential preconditions for lawful deployment.

Governing Legal Principles

Privacy Law

Government infringes on privacy when it forces people to reveal what they reasonably expect will be shielded from public view.[6] A minimum degree of privacy is guaranteed by the American constitutional requirement that the government respect citizens' "reasonable

expectation[s] of privacy" in certain protected zones, especially the home.[7] Congress and states have supplemented this baseline guarantee with statutory protections focused on particular kinds of information. Most applicable to Covid-19 are the health information privacy provisions of the Health Insurance Portability and Accountability Act (HIPAA), which restrict disclosures and uses of identifiable health information.

Governments may violate privacy when the volume of personal information intruded upon is excessive in relation to the government's purpose.[6] Governments may also violate privacy when they intrude into a space so intricately connected to a person's identity that the intrusion "depersonalize[s] and dehumanize[s]."[8] Health information privacy is commonly held to safeguard "personal dignity" and "protect[] patients from embarrassment, stigma, and discrimination."[9]

Antidiscrimination Law

American antidiscrimination law consists of two basic doctrines. First, disparate treatment (or intentional discrimination) occurs when an actor treats individuals differently because they are members of a protected class, such as a racial minority group.[10] When the actor engages in disparate treatment it must offer a justification, with the strength of the rationale calibrated to the protected class.

Second, disparate impact occurs when an actor takes a facially neutral action that differentially burdens a protected class. Disparate impact doctrine applies only in more limited circumstances outlined by statute, such as employment, healthcare institutions with federal funding, and housing.[10] In these domains, regulated parties may not use any tool that, *even unintentionally*, results in disparate outcomes, unless justified by "business necessity."[10]

It is straightforward to see how the use of AI to combat Covid-19 raises both antidiscrimination and privacy concerns. Data-hungry algorithms that calculate risk scores can pose privacy challenges because the personal information incorporated may be voluminous and divulge intimate personal information. Models commonly deploy gender and race, potentially running afoul of disparate treatment,[2,10] and risk scores may vary systematically across such demographic characteristics.

Evaluating Harms and Tradeoffs

Effectiveness

What legal standards must a government meet for deploying an ML application to fight Covid? In the context of a constitutional violation, policymakers must first show that an important government interest is at stake--an easy argument in the Covid-19 context. Next,

the government's action must be sufficiently well-tailored to serving that interest. In practice this distills to two questions: Is the policy likely to advance the government's objective (*effectiveness*)? And are there alternative ways of achieving that objective that are less burdensome on individual interests (*burden*)?

Although we highlight the constitutional analysis, quantifying the effectiveness and burdens of AI models is also relevant to assessing the legal permissibility of AI/ML models under principles of American administrative law (that is, the standards for actions of administrative agencies that legislatures set forth in statutes). For instance, courts can strike agency actions they conclude are "arbitrary and capricious." As applied to models deployed to fight Covid-19, this standard means that a health agency would need to provide a reasoned explanation for its use of the model and provide the evidence considered in its appraisal.

Courts and policy makers are often poorly equipped to assess the effectiveness and burdens of algorithmic tools, key aspects of which are summarized in Box 1. The problem is not merely a lack of technical competence. It is that courts and policymakers seeking to assess model performance will have to wade into the AI/ML field's replication crisis, not dissimilar from the one roiling the biomedical sciences.[11] The lack of incentives for robust evaluation is exacerbated in ML by model complexity, data volume, computational demands. This has resulted in influential models performing much worse in practice than originally reported. For instance, an ML-based risk score for whether a Covid-19 patient requires rapid response was found to have "limited value to guide clinical decision-making" for the vast majority of patients—but only after it was deployed to over 100 U.S. hospitals.[12] Similarly, epidemiological models predicting policy outcomes—some of which are ML-based[13]—have also proven to produce unreliable forecasts of actual Covid infection rates—but only after they were used to set policy.[14] Given the challenges facing even researchers steeped in technical know-how, policymakers and courts evaluating risk scoring models face a daunting task.

As governments turn to increasingly invasive tools focused on individualized predictions, the potential to harm privacy and antidiscrimination rights will grow. Proper evaluation of a given model will be at the heart of whether new tools appropriately trade off effectiveness and burdens.

Privacy

How can the government show that a particular AI tool is legal despite posing risks to privacy? While information closer to the core of a person's identity is more dangerous for the government to possess, even very intimate information is not sacrosanct. The government's interest may be so weighty that even the most personal information can be seized and

potentially used. For example, personal health information (PHI) is deemed highly sensitive under HIPAA and thus normatively bound up with dignitary concerns. But HIPAA allows information custodians to release PHI to prevent imminent risks to public health or safety. Similar logic underlies HIPAA's tolerance of limited PHI disclosures to public health agencies for surveillance and law enforcement.

Weighing these considerations is a complex endeavor. First, protecting privacy may degrade AI/ML models' accuracy. Second, in circumstances as severe as a pandemic, failure to deploy effective AI/ML tools can itself harm dignity. Privacy protections may then pose not just an accuracy-burden tradeoff, but a tradeoff *between* burdens:[15] sheltering for extended periods of time due to an ineffective tool may cause greater indignity than having PHI revealed. Third, privacy protections can themselves have disparate impact, degrading accuracy more for minority groups than majority groups.[16] At bottom, the government faces a highly complex tradeoff between effectiveness, dignity, *and* equality.

The leading technical framework for navigating these competing concerns is differential privacy, an approach that adds random noise to aggregate statistics computed from individual-level data. In this framework, policymakers can directly select the degree to which any individual's data influences aggregate statistics via a parameter, enabling policymakers to encode the extent of privacy protection. The technical fix, however, should not obscure the need for trade-offs. Policymakers may need to offer as much justification for *sacrificing* privacy as for *prioritizing* it.

Bias

Assessment of bias for Covid-19 risk scoring may seem more straightforward. Disparate treatment may well prohibit the use of a protected attribute (e.g., race) to generate a person's risk score. Disparate impact may occur when (a) outcomes (e.g., contracting or dying from Covid) are correlated with race *conditional* on the model's estimated risk score, or (b) when protected groups are not represented proportionally in different estimated risk categories.

But implementing these divergent metrics simultaneously is another matter entirely, raising profound questions of structural sources of bias. The risk of contracting Covid-19 appears linked to membership in protected groups and there appear to be significant differences in morbidity between African-American and Caucasian patients.[5] This creates a Catch-22. A model that is entirely blind to protected attributes, like race, may be more likely to produce risk scores correlated with those protected attributes. Deploying such a model to, say, determine which employees could return to work could well violate disparate impact. Technical solutions suggest adjusting the ML process to conform to fairness constraints, such as calibration parity across groups (i.e., that conditional on a risk score, outcomes are

independent of group status).[17] But calibrating risk scores by race raises significant constitutional concerns, as government classifications based on race are deemed particularly noxious. Moreover, various fairness constraints are often mutually incompatible.

It is possible that collecting a much wider range of socioeconomic predictors would eliminate the need for race variables in models. But collecting more such data may be infeasible or aggravate concerns about privacy by increasing the volume of data collected. Just as with differential privacy then, the pure engineering solution of imposing one fairness definition, given conflicting effects, cannot solve the underlying value tradeoffs.

The high likelihood of some differential impact of any modeling approach makes it all the more critical for policymakers to insist on reliable evidence as to the efficacy of proposed models. Knowing that ML tools may well engage in disparate treatment or cause disparate impact means that policymakers must be prepared to show that such tools are necessary to achieve public health goals, and to quantitatively establish the difference in efficacy between models that impose potential discrimination harms.

Toward An Evaluation Framework

The deployment of AI in the fight against Covid-19 is an important moment for algorithmic governance. There is an abundance of models, and a shortage of coordinated and consistent standards and evaluation. To give government uses of AI/ML the strongest prospects of passing legal muster, we spell out elements of a robust evaluation framework that addresses *effectiveness* and *burdens*.

First, the framework must be transparent to provide a basis for evaluation: source code, learned parameters, and base data should be released to the extent allowed by privacy concerns. Second, given the highly dynamic innovation space of AI, review should be distributed and decentralized. Third, evaluation methods and metrics must be thorough, robust, and interoperable, addressing performance across demographic groups and privacy-fairness tradeoffs. Last, interoperability permits evaluation results to be compiled in a single location, enabling decision makers to efficiently assess models.

One model framework is the National Institute of Standards and Technology (NIST) evaluation of facial recognition technology.[18] NIST enables any algorithm to be submitted and tested for accuracy, bias, and a range of other criteria using standardized tests and benchmark data. Specific methods for validating Covid-19 models have been examined by others,[12,14] but have fallen short on assessment of burdens. To be sure, the NIST framework is not perfect, particularly as facial recognition is deployed to a much wider range of domains not represented in the NIST data. But NIST provides a good example of the kind of robust evaluative framework that the law may ultimately demand. After development and evaluation

through this framework, deployments can then also be evaluated through adaptive trials to assess operational performance.[19,20]

Robust evaluation will not only help AI applications survive legal challenges, but also cultivate public trust in a highly contentious time for AI governance and public health.

References

1. Mozur P, Zhong R, Krolik A. In Coronavirus Fight, China Gives Citizens a Color Code, With Red Flags. The New York Times [Internet]. 2020 Mar 1 [cited 2020 May 18]; Available from: https://www.nytimes.com/2020/03/01/business/china-coronavirus-surveillance.html
2. DeCaprio D, Gartner J, Burgess T, Kothari S, Sayed S. Building a COVID-19 Vulnerability Index. ArXiv Prepr [Internet]. 2020; Available from: https://arxiv.org/abs/2003.07347
3. Carnegie Mellon University. COVID Voice Detector [Internet]. Available from: https://cvd.lti.cmu.edu/
4. Burt C. Facial recognition and fever detection products launched by Remark Holdings, MSP introduces biometric kiosk [Internet]. Biometric Update. 2020 [cited 2020 Jun 9]. Available from: https://www.biometricupdate.com/202005/facial-recognition-and-fever-detection-products-launched-by-remark-holdings-msp-introduces-biometric-kiosk
5. Yancy CW. COVID-19 and African Americans. JAMA. 2020 May 19;323(19):1891–2.
6. Carpenter v. United States. Vol. 585, U.S. 2018. p. ____.
7. Katz v. United States. Vol. 389, U.S. 1967. p. 347.
8. Sklansky DA. Too Much Information: How Not to Think about Privacy and the Fourth Amendment. Calif Law Rev. 2014;102(5):1069–122.
9. Gostin LO, Nass S. Reforming the HIPAA Privacy Rule: Safeguarding Privacy and Promoting Research. JAMA. 2009 Apr 1;301(13):1373–5.
10. Ricci v. DeStefano. Vol. 557, U.S. 2009. p. 557.
11. Hutson M. Eye-catching advances in some AI fields are not real [Internet]. Science | AAAS. 2020 [cited 2020 Jun 3]. Available from: https://www.sciencemag.org/news/2020/05/eye-catching-advances-some-ai-fields-are-not-real
12. Singh K, Valley TS, Tang S, Li BY, Kamran F, Sjoding MW, et al. Validating a Widely Implemented Deterioration Index Model Among Hospitalized COVID-19 Patients [Internet]. Health Informatics; 2020 Apr [cited 2020 May 25]. Available from: http://medrxiv.org/lookup/doi/10.1101/2020.04.24.20079012
13. Dandekar R, Barbastathis G. Quantifying the effect of quarantine control in Covid-19 infectious spread using machine learning. medRxiv. 2020 Jan 1;2020.04.03.20052084.
14. Marchant R, Samia NI, Rosen O, Tanner MA, Cripps S. Learning as We Go: An Examination of the Statistical Accuracy of COVID19 Daily Death Count Predictions. ArXiv Prepr [Internet]. 2020 May 3 [cited 2020 May 18]; Available from: http://arxiv.org/abs/2004.04734
15. Ruggles S, Fitch C, Magnuson D, Schroeder J. Differential Privacy and Census Data: Implications for Social and Economic Research. AEA Pap Proc. 2019 May 1;109:403–8.
16. Bagdasaryan E, Poursaeed O, Shmatikov V. Differential Privacy Has Disparate Impact on Model Accuracy. In: Advances in Neural Information Processing Systems 32. 2019. p. 15479–15488.
17. Barocas S, Hardt M, Narayanan A. Fairness in Machine Learning [Internet]. 2019. Available from: https://fairmlbook.org/
18. Grother P, Ngan M, Hanaoka K. Face Recognition Vendor Test (FRVT) Part 2: Identification [Internet]. (NIST Interagency/Internal Report (NISTIR)). Report No.: 8271.

Available from: https://doi.org/10.6028/NIST.IR.8271
19. Horwitz LI, Kuznetsova M, Jones SA. Creating a Learning Health System through Rapid-Cycle, Randomized Testing. N Engl J Med. 2019 Sep 19;381(12):1175–9.
20. COVID + AI: The Road Ahead [Internet]. Stanford HAI. [cited 2020 Jun 3]. Available from: https://hai.stanford.edu/watch-covid-ai-road-ahead

Box 1: Main dimensions of effectiveness and burdens of AI/ML systems

	Issue	Example of failure
Effectiveness	Accuracy	Hospitalization risk model fails to predict actual hospitalizations
	Replicability	Feature selection for risk model cannot be replicated
	Generalizability	Risk models works in one hospital, but not another
	Explainability	Outputs of the risk model cannot be explained easily to a human user, limiting takeup rate by decision makers
Burden	Bias	Risk model performs well only on one demographic group on which it is trained
	Privacy	Risk model uses and/or discloses sensitive information about individual
	Due care / process	Individualized patient assessment is compromised due to risk score