Revision Memorandum

To:	Dr. Diana Lucifero
From:	Mark Krass, Peter Henderson, Michelle M. Mello, David M. Studdert, and Daniel E. Ho
Re:	Decision on Ms. No. BMJ-2020-059371 ("Artificial Intelligence, Covid-19 and Law: The Need for Evaluation in the Age of Many Models")
Date:	August 26, 2020

Dear Dr. Lucifero,

We are thankful for the helpful reviews and suggestions.

We have attached a manuscript that responds to the very helpful comments and have enclosed point-by-point replies below. The original comments are set out in *italics* for reference and our responses are indented.

Many thanks for your work on the article. We believe the manuscript is much strengthened as a result of this feedback and we hope you agree.

We'd look forward to seeing the work appear in *BMJ*!

MK, PH, MMM, DMS, DEH

Point-by-Point Reply to Editors' Comments

The editors would like to acknowledge the commendable efforts made in the preparation of this manuscript. We thought that this was an interesting article and we hope that these comments will be helpful to the authors.

We thank the editors and the reviewers for the detailed feedback on the manuscript and on the positive reception! We're thrilled that you enjoyed the article and we have found the comments quite helpful to refine the exposition.

Please add missing author positions.

We added author positions.

Please revise the title to reflect that the focus of the article is the US.

We have included the word 'U.S.' in the title and in the Standfirst abstract to alert readers to the focus of our piece.

A clinical colleague suggested that the structure of the article should be reconsidered and perhaps flipped around given that the focus on impacts to individuals and communities would be of greater interest to our readers than those on courts and policymakers.

We appreciate the clinical colleague's perspective to provide more of a focus on the impact on individuals and communities. First, we have added a running example of risk scores used to allocate health care resources for at-risk and infected individuals at the Department of Veterans Affairs (VA), which serves roughly 9 million individuals annually. We believe this running example, which is added in the introduction, underscores the substantive magnitude of these decisions. Second, in response to feedback by Reviewer 1 and 2, we have developed a new passage in the last section of the article that provides much more concrete detail on the application of the evaluation framework to that VA scoring system.

After adding that passage, however, we deliberated extensively about whether to flip the structure of the piece. For several reasons, we have opted against flipping the organization. While we agree that the impact on individuals is important, we understood the contribution of the piece to be primarily about how legal constraints require greater care with evaluation of AI-based solutions. It would hence appear awkward to have either (a) the application of legal principles, or (b) the evaluation framework appear before providing the more general landscape. If the editors have other suggestions on how to reorganize, we would welcome them, but we have also given extensive thought for how this piece would most effectively communicate the ideas to a general audience that may not have much background to law, let alone U.S. law. Nonetheless, we thank the clinical colleague's perspective for inspiring us to make the more moderate changes outlined above.

Reviewer 3 is a patient reviewer and we have contacted her to try and get some clarity about her comments.

We are grateful for reviewer 3's comments. As we detail below, the references provided by Reviewer 3 discuss potential privacy breaches by Facebook – e.g., the suicide surveillance system, the solicitation of health information, and the lack of attention to health data in the Facebook settlement with the Federal Trade Commission. We agree that these are fascinating examples of some of the challenges at the intersection of technology, health, and privacy.

While those applications are not about COVID-19, and hence outside the scope of our article, we agree with the referee that a more tangible example can help anchor the analysis. We have hence revised the manuscript to use a running example of an AI-based system at the VA Hospitals, which we use at the end to illustrate more concretely the evaluation framework.

Again, we thank the editors for their time and work on this manuscript. We believe the revision has given us the chance to substantially strengthen the exposition, and we hope you agree!

Point-by-Point Reply to Comments by Reviewer 1

This piece addresses an interesting area: the legality of AI-powered tools used by governments in the response to the COVID-19 pandemic. It raises the issues of privacy and bias, notes potential concerns, and presents a very brief framework for evaluation. While the piece is intriguing, several areas of concern stood out.

> We are thankful for Reviewer 1's comments and delighted that the reviewer found the contribution to be intriguing.

It is unclear what government uses, specifically, the authors view as potentially concerning. The uses they cite all appear to be either research uses, products in commercial development, or, in one instance, a use by the Chinese government. Are there proposals for U.S. government use of AI related to COVID-19 with real consequences (e.g., denial of benefits, denial of care, or the imposition of sanctions)? If so, the authors should be clear about those examples. If not, the authors should indicate why they think such uses might be plausibly on the horizon.

> We have revised the examples we use to be specific to uses within real public sector settings. For example, the Veterans' Affairs Department uses AI for risk scoring according to recent sources and prisons search for Covid symptoms in the transcripts of inmate phone calls. While our running example is now grounded in a documented instance of an agency conditioning certain benefits (e.g., hospital admission) on the outcomes of AI/ML algorithms, we also note that many of the research-phase tools are prominent enough to warrant serious analysis, even if they have yet to be implemented. We thank the reviewer for helping us tighten the focus of the motivating examples.

There is a tension between the article's theoretically international focus—the first example of AI COVID-19 software is from China, and the second paragraph references "governments"—and the reference to U.S. privacy law only, which is vastly different from Chinese privacy law and substantially different from EU, which notably has the GDPR rather than the substantially restrained HIPAA Privacy Rule or the Fourth Amendment. Similarly, the only antidiscrimination law mentioned is American. Certainly, a survey of international laws is of broad scope, but the authors should be clear that the analysis really only applies in the U.S.

> This is an excellent point. While the topics and tradeoffs we discuss may generalize to other contexts, the specifics will likely not. We have revised the scope of our title and subsequent discussion to be specific to U.S. law. We have made this clear in the title, the Standfirst, as well as the introduction. In addition, we have foregrounded the discussion of HIPAA in the section on privacy principles.

It would be useful for the authors to justify why Fourth Amendment jurisprudence is the right frame to evaluate any government acquisition of interest. Is public health surveillance is typically regarded as a search? Making this point explicit would be helpful.

> We appreciate the reviewer's pointing out that the Fourth Amendment has circumscribed applicability in the public health setting. We have shifted our summary of U.S. privacy law such that the Fourth Amendment is less prominent. We draw from the theoretical

framing on the Fourth Amendment not because a Fourth Amendment challenge is necessarily forthcoming, but rather because the major theoretical treatments of privacy in U.S. law emanate from constitutional analysis. Because we are writing at a general level, we use some of these frameworks to inform our analysis. We nonetheless more quickly pivot to the related concerns under HIPAA.

Are the authors suggesting that the use of AI tools in a pandemic for public health purposes has any realistic chance of failing rational basis review? Or are they suggesting some other level of scrutiny is appropriate? The legal analysis is hazy.

We appreciate this question. We have clarified that the standard of review would depend on the cause of action and the particular nature of the harm. We now emphasize our main argument: Across multiple potential causes of action and multiple standards of review, the basic evidentiary question of how to measure benefits and burdens will remain, and courts ought to demand high-quality evidence along those lines.

Who is doing the evaluating? Do these AI uses go through FDA or something else? The authors suggest in a single line that "review should be distributed and decentralized," but what does that mean? In general, the proposed framework for evaluation is extraordinarily brief. This is unfortunate, because the framework is sensible, and the concerns raised are real; there are significant tradeoffs involved. The article would be better served with less text devoted to unlikely-to-materialize constitutional challenges and more text addressed to the underlying policy tensions and a more fleshed-out set of solutions for addressing them.

We are grateful for Reviewer 1 for pushing us to spell out more concretely how our evaluation framework might be implemented. We've expanded the final section of the paper (cutting much language in the earlier sections to remain under the 2,000 word limit!). We clarify that the entity responsible for administering the evaluation need not be any particular agency, but should ideally be *independent* of model development. We suggest NIST may be a possible option, but a group within an agency employing a particular model, or an academic clearinghouse could also run such an evaluation. To flesh out this out concretely, per the reviewer's suggestion, we use Veterans Affairs (VA) as a real world example of an agency using a risk assessment model in VA hospitals. We suggest that either the VA, NIST, or some other independent third party might be responsible for evaluation. We have also clarified that under multiple tiers of scrutiny, independent evidence of efficacy might be relevant.

Finally, the piece has some strange references. It cites multiple U.S. Supreme Court cases that are only moderately on point, but essentially no legal scholarship on, e.g., algorithmic governance, medical AI, health privacy, or discrimination. Prince and Schwarcz on proxy discrimination by AI (2020) is particularly on point.

We appreciate the reviewer's suggestion to expand the set of citations. In contrast to law review articles, the total number of references is limited to 20, so we have wrestled with what might be the best materials are to cite (but would be happy to add more if the number of references could be expanded above 20). We have taken out one of the case citations, but opted to leave in some of the principal Supreme Court cases, as those strike us as the primary drivers of any legal analysis. We recognize that there is a voluminous literature on fairness and algorithmic governance, which is why we cite to one of the leading textbooks on the topic:

Barocas S, Hardt M, Narayanan A., Fairness in Machine Learning. That reference directs the reader to many secondary references and has an extensive treatment of topics like "proxy discrimination" (the topic of Prince and Schwarcz), showing that protected attributes can be inferred with probability approaching one as the feature set grows. We thank the referee for pointing out the connection to the secondary literature and we couldn't agree more about its importance!

We thank Reviewer 1 for the thoughtful feedback that has given us the chance to strengthen the manuscript substantially.

Point-by-Point Reply to Comments by Reviewer 2

This is a well-written paper on a timely and important topic. It captures well the key issues related to US privacy and discrimination law raised by the use of AI models in the context of COVID19. The analysis is also broadly relevant for other jurisdictions. This is a short commentary, and many issues could obviously be more developed, but the purpose of the paper is to sketch the general framework under which the use of AI should be assessed, focusing particularly on issues of effectiveness, and the paper succeeds in that. The question whether the privacy invasions related to AI are not burdening more than necessary is not really discussed, but that seems to be outside of the scope of the paper (and is a very complex question in the context of the uncertainty surrounding COVID19).

> We thank the reviewer for the enthusiasm about the article! We are grateful that the reviewer appreciates the broad relevance for other jurisdictions and that the reviewer believes the piece succeeds in its aims of articulating the general framework.
>
> We also agree with the reviewer that it would be ideal to flesh out the arguments about whether privacy invasions burden more than necessary. The length restriction does make it difficult to address those complex questions, as the referee alludes to. We have, as a result, focused on the core argument that the legal risks in AI for COVID-19 require more robust evaluation of performance.

The only substantial comment I have is that the final recommendations could benefit from some further elaboration: the NIST model of evaluation of facial recognition technology is recommended, but summarized in one sentence. Is more to be said about this model? I would also suggest to add to the recommendations the need for fully independent evaluation. This speaks for itself, and I know the authors themselves subscribe to this, considering their work on these issues. But in light of the conflicts of interest that have become apparent again in publications of COVID19-related studies, it is worth repeating.

> We thank the referee for this suggestion. In response, we have expanded substantially our discussion to emphasize the need for independent evaluation. The last section adds the following paragraph:
>
> > To illustrate, consider the application to the VA's adoption of a risk scoring model for Covid-19 hospitalized patients. Under the framework laid out above, competing vendors would submit their models in a standardized format to an independent party, such as NIST or an academic clearinghouse. The third party would run each vendor's tool on a hold-out set of data, providing an authoritative audit of the benefits and burdens they offer. For example, the agency might audit the degree to which a given tool's performance depends on access to invasive data or the extent to which it scores protected sub-groups differently. Ideally, the evaluator's process would be a stable, interoperable pipeline—much like NIST's facial recognition evaluator—such that the assessment process is not resource-intensive.
>
> We hope that this additional detail provides more detail and also emphasizes, per Reviewer 2, the need for *independent* evaluation.

p. 3: L57: government "infringes on" : I read 'infringes' as referring to a violation. But there are instances where the disclosure of information over which one has a reasonable expectation of privacy is legitimate. So better 'intrudes upon'? Or is this the language used in the case?

> We've adjusted our text to reflect the reviewer's helpful suggestion, using the word "intrudes" instead.

p. 4 Line 11: Governments 'may violate': doesn't the term 'excessive' indicate a situation where there is a violation of privacy?

Line 12: 'excessive': I'm not so familiar with the terminology used in the US privacy statutes, but is 'disproportionate' not a better term? Perhaps this is why the term 'may' is used: excessive may not be disproportionate if the goal of information gathering is to deal with an extraordinary risk

> We've adjusted our text to reflect the reviewer's helpful suggestion by changing the word "excessive" to "disproportionate."

Line 58-59: "In the context of a constitutional violation": perhaps better 'In the context of an allegation of' or 'to decide whether there is a "constitutional violation" [not so clear also why 'constitutional' is used here. Perhaps better: In the context of an allegation of violation of fundamental rights"... This may be jurisdictional difference in phrasing.

> This is a fair point. We have edited the language to clarify that the claim involves an allegation, not a violation per se.

p. 5: line 57: 'more dangerous to possess': the term dangerous seems odd here. Authors mean that it is more likely to be considered a violation. Alternative: 'problematic' [if 'more likely to be considered a violation... ' is too long]

> We appreciate this careful attention to the terminology! We've adjusted our text to change "dangerous" to "problematic" per the reviewer's helpful suggestion.

p. 6: line 14: 'failure to deploy AI can itself harm dignity". Odd phrasing: the failure not to use a technology is in and of itself not causing a dignitary harm. It indirectly may lead to it, in the authors' opinion, when it results in a failure to deal with pandemic control differently than through other more intrusive measures.

> We've adjusted our text to make this more clear that we in fact were referring to the harms caused by using less effective or overly broad alternatives.

Line 43: I don't understand well point a: can this be stated differently? (after 'conditional on... ' is the issue that may not be so clear)

> We thank the reviewer for flagging this ambiguity. We have clarified the sentence to read: "Disparate impact may occur when the use of the risk score leads to decisions (e.g., preventing someone from going to work) that affect racial groups differentially."

Again, we are grateful for Reviewer 2's enthusiasm for the manuscript and the careful attention to phrasing!

Point-by-Point Reply to Comments by Reviewer 3

My main feedback is there needs to be more tangible examples cited, rather than what may happen under the Privacy Law section.I recommend reviewing the following sources:
https://www.acpjournals.org/doi/10.7326/M19-0366
https://missingconsent.org/facebook-patient-ftc-complaints/
https://www.statnews.com/2019/07/31/facebook-ftc-settlement-health-information-privacy/

We thank Reviewer 3 for this feedback as well as these additional references. These articles each appear to discuss distinct potential privacy breaches by Facebook – e.g., the suicide surveillance system, the solicitation of health information, and the lack of attention to health data in the Facebook settlement with the Federal Trade Commission – which we agree are a fascinating examples of some of the challenges at the intersection of technology, health, and privacy. While those applications are not about COVID-19, and hence outside of the scope of our piece, we agree with the reviewer that a more tangible example can help anchor the analysis. We have hence revised the manuscript to use a running example of an AI-based system at the VA Hospitals, which we use at the end to illustrate more concretely the evaluation framework.

We thank Reviewer 3 for the feedback!