Canadian Journal of Bioethics
Revue canadienne de bioéthique

Canadian Journal of Bioethics
Revue Canadienne de Bioéthique

Commentaire critique / Critical commentary

AI Bias in Healthcare: Using *ImpactPro* as a Case Study for Healthcare Practitioners' Duties to Engage in Anti-Bias Measures

Samantha Lynne Sargent

Volume 4, numéro 1, 2021

URI : https://id.erudit.org/iderudit/1077639ar
DOI : https://doi.org/10.7202/1077639ar

Aller au sommaire du numéro

Éditeur(s)

Programmes de bioéthique, École de santé publique de l'Université de Montréal

ISSN

2561-4665 (numérique)

Découvrir la revue

Citer ce document

Sargent, S. L. (2021). AI Bias in Healthcare: Using *ImpactPro* as a Case Study for Healthcare Practitioners' Duties to Engage in Anti-Bias Measures. *Canadian Journal of Bioethics / Revue canadienne de bioéthique*, *4*(1), 112–116. https://doi.org/10.7202/1077639ar

Résumé de l'article

L'introduction d'*ImpactPro* pour identifier les patients ayant des besoins de santé complexes suggère que les préjugés actuels et les impacts des préjugés dans les IA de soins de santé proviennent de pratiques historiquement biaisées menant à des ensembles de données biaisés, d'un manque de supervision, ainsi que de préjugés chez les praticiens qui supervisent les IA. Afin d'améliorer ces résultats, les praticiens de la santé doivent adopter les meilleures pratiques actuelles en matière de formation à la lutte contre les préjugés.

Canadian Journal of Bioethics
Revue Canadienne de Bioéthique

SL Sargent. *Can J Bioeth / Rev Can Bioeth.* 2021;4(1):112-116

COMMENTAIRE CRITIQUE / CRITICAL COMMENTARY (ÉVALUÉ PAR LES PAIRS / PEER-REVIEWED)

AI Bias in Healthcare: Using *ImpactPro* as a Case Study for Healthcare Practitioners' Duties to Engage in Anti-Bias Measures

Canadian Bioethics Society / Société canadienne de bioéthique
Concours de rédaction / Writing Contest

Samantha Lynne Sargent[a]

Résumé
L'introduction d'*ImpactPro* pour identifier les patients ayant des besoins de santé complexes suggère que les préjugés actuels et les impacts des préjugés dans les IA de soins de santé proviennent de pratiques historiquement biaisées menant à des ensembles de données biaisés, d'un manque de supervision, ainsi que de préjugés chez les praticiens qui supervisent les IA. Afin d'améliorer ces résultats, les praticiens de la santé doivent adopter les meilleures pratiques actuelles en matière de formation à la lutte contre les préjugés.

Mots-clés
intelligence artificielle, apprentissage automatique, préjugés, préjugés implicites, racisme, ImpactPro

Abstract
The introduction of *ImpactPro* to identify patients with complex health needs suggests that current bias and impacts of bias in healthcare AIs stem from historically biased practices leading to biased datasets, a lack of oversight, as well as bias in practitioners who are overseeing AIs. In order to improve these outcomes, healthcare practitioners need to engage in current best practices for anti-bias training.

Keywords
artificial intelligence, machine learning, bias, implicit bias, racism, ImpactPro

Affiliations
[a] Department of Philosophy, University of Waterloo, Waterloo, Canada
Correspondance / Correspondence: Samantha Lynne Sargent, slsargen@uwaterloo.ca

INTRODUCTION

Artificial Intelligence (AI) and other Machine Learning (ML) applications are increasingly permeating all aspects of our lives, and healthcare is no exception. In addition to emerging applications, AI is already used in a variety of ways including medical imaging, parsing and collating electronic medical records, optimizing care trajectories, diagnosing, improving enrollment in clinical trials, and even in reducing medical errors (1-4). This is not an exhaustive list; suffice it to say that the applications are as varied and complex as the medical field itself. In 2018, the Nuffield Council of Bioethics flagged potential issues in the use of AI in healthcare due to the way in which AI reproduces biases in data sets that are used to train ML algorithms, and also the ways in which biases can be "embedded in the algorithms themselves, reflecting the beliefs and prejudices of AI developers" (2). In this paper, I take bias to refer to both consciously and unconsciously held negative feelings or opinions towards marginalized groups, rooted in historic discrimination, that influence the way a person behaves and thinks.

These biases and their negative effects on health are already being seen in cases such as the recent *ImpactPro* study, which found that an algorithm in New York's United Health Services failed to recommend Black patients to a complex health needs program at the same rate as White patients (5-6). Therefore, it is imperative that the healthcare field respond to the proliferation of such technologies in order to correct previous inequities in the healthcare system that have produced the biased data that AI technologies are currently reproducing (4,7). To do so, healthcare practitioners must engage in a variety of anti-bias measures, such as implicit bias training, education on bias in medicine, and "perspective taking", as well as assuming their responsibilities as overseers of and collaborators with AI technologies. Many measures that are currently used to reduce bias in everyday healthcare interactions can be transferred to use with AI, especially when healthcare practitioners have the final say in decisions that ML algorithms recommend. It is hard to determine universal rules for AI applications in healthcare because the applications, uses, and contexts are so incredibly diverse, and evolving all the time. Given this, I will use the *ImpactPro* case to illustrate the ways in which the effects of AI on healthcare reaffirm existing duties to combat bias when delivering care to better serve the health needs of marginalized patients. I argue that the *ImpactPro* case shows that there are opportunities for healthcare practitioners to push back against bias in AI algorithms by decreasing biased practices within hospitals and medical research, and by building trust with marginalized communities, with the eventual goal of bettering the data with which AI is trained, and more quickly catching cases where AI results are biased. These avenues are also in alignment with principles for AI best practices, such as those advanced by the Montreal Declaration for the Responsible Development of Artificial Intelligence and the High Level Expert Group (HLEG) on Artificial Intelligence.

THE *IMPACTPRO* STUDY

An October 2019 study showed that *ImpactPro*, a healthcare algorithm which was supposed to identify patients with complex health needs who would benefit from greater attention and additional resources, was biased against recommending Black

ISSN 2561-4665

patients to the complex needs program. Although the main goal of using the algorithm was to better serve patients, an additional goal was also “in part” to reduce costs (6). As a result, the algorithm was formulated in such a way that it used healthcare costs as a proxy for healthcare needs, i.e., it predicted complex needs by predicting which patients would have the greatest future healthcare costs. This makes a certain kind of sense, as healthcare costs are easily quantifiable, whereas “health” itself is a more nebulous and harder to measure concept (6). The algorithm was fed data regarding insurance claims, insurance types, medications, etc. but not race. No mention is made of other identity categories (such as gender) being used or not, either by the initial algorithm or the researchers’ reformed algorithm (6). The researchers investigating the algorithm found that in regards to race, Black patients generated different healthcare costs than White patients, such as emergency costs versus surgical and outpatient specialist costs. However, this discrepancy can be explained through larger societal factors, particularly in the US where poor patients face systemic barriers when it comes to accessing healthcare even if they have insurance (6). Similar issues have been found in AI systems that used postal codes as predictors of the length of hospital stays (8). Since race is largely correlated with socio-economic income and status, Black patients face further barriers in access to care, which in turn affects their ability to spend money on healthcare (8). There is also evidence that when Black patients have White family doctors, these patients are less likely to participate in recommended preventative care because of historic factors that reduce Black patients’ trust in the healthcare system, and because White healthcare providers view Black patients as having different levels of intelligence, pain tolerance, and feelings of affiliations with patients, which is thought to result in different communication styles, further erosion of trust, and the introduction of bias (6,9).

All of these factors combined mean that Black patients tend to spend less on their healthcare, and so the algorithm did not identify Black patients being as sick as White patients because it believed (perhaps rightly so) that they were going to spend less money on their care. In a ML system, this is called a “labelling bias,” and can be corrected by changing the label on which the algorithm works or by re-weighing data sets (10). Since this bias arose from an initial design problem in the algorithm due to using an inaccurate quantifiable measure (healthcare spending) as a proxy for complex healthcare needs, the study authors suggested that fixing the label would be fairly straightforward and so they produced a second algorithm which still excluded race data but changed the label to include a quantified criterion of health (6). They noted that picking the correct label requires “deep understanding of the domain, the ability to identify and extract relevant data sets, and the capacity to iterate and experiment,” (6). Other research shows that label biases can be fixed by examining and re-weighing data sets without changing labels, although this remains theoretical work, and oversight is still necessary to uncover labelling biases in the first place (10).

What is also notable about this study is that the initial algorithm was not simply making decisions in a vacuum as to whether or not to enroll patients in the complex health needs program. The algorithm automatically identified patients in the 97th percentile for enrollment, but anyone over the 55th percentile was referred to their primary healthcare provider to make the decision (6). Those healthcare providers were given electronic health records and insurance claims, and did positively correct for the bias of the initial *ImpactPro* algorithm. However, the providers still demonstrated bias against their patients in comparison to the *ImpactPro* study researchers’ reformulated algorithm which identified patients for enrollment based on a quantification of health. The healthcare practitioners were not as biased as the initial *ImpactPro* algorithm, but were still biased against Black patients, and were still more biased than the revised algorithm. While the healthcare practitioners played no role in fixing the algorithm, they did have the opportunity, prior to the researchers engaging with *ImpactPro*, to fix its outcomes by being more attentive to the algorithm’s assessments and less biased in their recommendations of patients to the program. The phenomenon of algorithms being biased should be unsurprising to healthcare practitioners given that the insurance data that made the algorithm biased in the first place came from healthcare providers, and the healthcare system’s own biases. How can we expect the humans who created the error in the first place to fix it without some larger structural changes?

BIAS IN HEALTHCARE

This example succinctly illustrates the true responsibility when it comes to healthcare providers’ promotion of equity both when they use AI technologies, and when they do not. It is unclear if it will ever be possible to eradicate bias in individuals, or in AI systems (11). Nonetheless, healthcare providers are part of the ecosystem that creates the data that the ML algorithms are fed, and so healthcare providers should have the opportunity to scrutinize and question individual outputs of AI systems that they use. The main duty that healthcare providers should have in regard to promoting equity when using AI systems is to reduce bias in their own care practices, and to advocate for various educational projects that inform healthcare practitioners about the negative effects of implicit bias in medicine. Furthermore, healthcare practitioners are well positioned to intervene on AI-related issues, as while there are similar ethical concerns in both healthcare and AI development, there is a more robust history of ethics and professional norms in health professions, however imperfectly applied (12). These solutions require no technical acumen, and are not unique to AI issues.

It is well documented that even in the absence of AI, healthcare practitioners are particularly susceptible to influences of implicit bias because of the way that diagnostic processes may “promote reliance on stereotypes for efficient decision making,” the ways in which physician training “emphasizes group level information,” and because their “vast knowledge of scientific data may create a strong belief in their personal objectivity, promoting bias in decision-making” (13-14). In general, it is demonstrated that physicians in the West have strong pro-White biases, and both implicit and explicit biases have also been shown regarding criteria such as obesity, gender, and age, in addition to race (13-14). Furthermore, there are demonstrated links between implicit bias and disparate treatment decisions for minorities (9,13-14). While *ImpactPro* and many of these other studies rely on the American context, where anti-Black racism is primarily an issue, there is good evidence that in the Canadian

context, First Nations, Inuit, and Métis patients are also discriminated against and face bias when seeking healthcare. In emergency rooms, these patients face stereotypes that may lead to inadequte triage. Due to historic acts of discrimination in healthcare contexts, such as forced sterilization, First Nations, Inuit, and Métis patients may be similarly reluctant in seeking healthcare, as are Black patients in the US (15). It is likely that these biases will be reproduced through AI use if the ML algorithms are trained on Canadian healthcare datasets. Regardless, all of these biases and their resulting effects on equity in patient care exist without the use of and reliance on AI technologies.

BEST PRACTICES FOR THE ELIMINATION OF BIAS

One best practice to compensate for implicit bias in physicians is to increase awareness of susceptibility to implicit bias amongst physicians through education (14,16-17). This should include education on implicit bias, as well as education regarding the fact that the evidence that practitioners use to diagnose is also biased, given the historic failure of randomized trials to adequately identify diseases via standard symptoms and predict the efficacy of interventions and drugs when it comes to minority groups (7). Other strategies include “individuating,” which “involves conscious effort to focus on specific information about an individual, making it more salient in decision-making than the person’s social category information,” and “perspective-taking” which involves asking healthcare providers to imagine the feelings of patients (14). Finally, in addition to implementing these behavioural changes, biases and healthcare disparities could be significantly reduced by increasing diversity in the population of healthcare providers, specifically by increasing the number of African American/Black physicians, as members of that group have been found to demonstrate significantly less race bias (14). The relative merits of these strategies have not yet been compared, but given evidence of their efficacy, engaging in all these actions is a good first step (14). Therefore, healthcare providers and trainers of healthcare providers already have a duty to educate themselves regarding implicit bias and practice individuating and perspective taking strategies, especially when they are treating patients from marginalized groups, and are themselves not members of those groups.

When we look at these best practices regarding the mitigation of implicit bias in physicians’ work, we can see parallels as to how these strategies might work when using AI, especially in cases like *ImpactPro* where healthcare providers have the final say. For example, educating healthcare providers on the impact of implicit bias, and medicine’s history of bias, might be a good way to enable them to recognize when ML algorithms may be giving biased outputs. “Individuating” in regard to oversight also seems to be a very strong technique going forward. We can imagine that had the healthcare providers with oversight in the *ImpactPro* case individuated their patients and looked more closely at their individual situation, specific information, etc. they might have referred patients with complex health needs to the program at a level equal to the secondary AI system researchers created to adequately evaluate health needs. Similarly, “perspective-taking” on the part of physicians is something that AI systems are currently ill-equipped to do, but humans excel at these sorts of “social intelligence” tasks that require the recognition of human emotions (19). These steps are also in alignment with the ethical principles for AI best practices, including those put forward by the Montreal Declaration for the Responsible Development of Artificial Intelligence such as a concern for well-being, equity, diversity inclusion, prudence, solidarity, and responsibility. Anti-bias measures can ensure that the use of AIs “permit the growth of the well-being of all sentient beings,” “contribute to a just and equitable society” including a “more equitable and mutual distribution of individual and collective risk”, maintain “social and cultural diversity,” and anticipate “the potential adverse consequences of AIS use…by taking appropriate measures to avoid them”, and not diminish the “responsibility of human beings when decisions must be made” (20). Similarly, these steps are in alignment with HLEG’s AI Guidelines on trustworthy AI, as they emphasize human agency and oversight, transparency, diversity, non-discrimination and fairness, as well as accountability (21). Therefore, anti-bias training is in alignment with the goals of responsible AI development, and ensures that healthcare practitioners maintain responsibility for their care duties. While these AI protocols may have yet more stringent requirements for the development of AI, in this article I am primarily concerned with the ethical implementation and use of existing AI technologies. Higher level AI guidelines such as those proposed in the Montreal Declaration and the HLEG’s AI guidelines should be used when developing anti-bias training and creating datasets to ensure they will be applicable to AI use in such cases. Further, there is more work to be done in translating these high-level guidelines into clinical, educational, and research policies, as well as curriculums.

In addition to these steps, when healthcare practitioners using AI suspect or are unsure of biases in their software, easy steps exist to have that software investigated. The Algorithmic Justice league currently accepts submissions of all types of ML software to be investigated for bias when it is suspected (19). They also accept submissions of technology for bias checks on the part of technology producers. When using AI software in the healthcare system, these sorts of checks and oversights prior to implementation should be seen as routine as the checks and oversights on new treatment methods, and their implementation should be continuously monitored for bias, for which bias training should help. Finally, increasing diversity in regard to who is working in healthcare, who is working in AI, and who is being included in the data sets that ML algorithms are being fed, are all ways forward to identify problems, decrease bias and increase equity in this field and in healthcare outcomes. Clearly some ethical issues remain when it comes to the improvement of AI, such as how to better increase the quality of health data overall without violating privacy or placing undue burdens on minority populations to participate in research (8). One example of such burdens are those that occurred in the Henrietta Lacks and Tuskegee cases where Black populations were exploited for questionable research benefits. However, my proposals focus on concrete first steps for frontline AI users who also interact directly with patients. Larger questions of improvement and regulation are often beyond the capacity and knowledge realms of front-line healthcare providers (22). Despite this, it is possible that through increased bias awareness, especially for physicians or other healthcare practitioners who have control over patient diagnoses and referrals, greater trust in the healthcare system may be achieved for minority groups, and this may then lead to better and more inclusive data sets.

CONCLUSION

The biggest potential pitfall with AI is seeing it as a solution to our very human faults, rather than as a tool that reflects what we have done in the past. Just as we need to improve ourselves going forward, we need to also improve AI outputs. The only way that this is going to happen is by first improving ourselves. The challenges posed by AI in healthcare settings are not new challenges. If we start viewing it not as a separate duty to promote equity in the implementation of AI systems, but rather as a part of an existing duty to promote equity and reduce bias in all aspects of care, that is when we will see better healthcare outcomes and better treatment of the most vulnerably socially situated patients.

Reçu/Received: 24/04/2020

Publié/Published: 01/06/2021

Remerciements
Merci au département de philosophie de Waterloo et en particulier à Katy Fulfer pour son aide et la révision de ce manuscrit; à mon réviseur Carl Mörch pour ses suggestions; à Amanda et Sydney, les étudiants de la Société canadienne de bioéthique et les administrateurs du concours de rédaction du SCB-RCB; à mes réviseurs anonymes; et à Nathalie Brown pour ses commentaires et suggestions.

Acknowledgements
Thank you to the Waterloo Philosophy Department and in particular Katy Fulfer for her review and help with this paper; to my peer-reviewer Carl Mörch for his suggestions; to Amanda and Sydney, the students at large for the Canadian Bioethics Society, and administrators of the CBS-CJB student essay contest; to my anonymous reviewers; and to Nathalie Brown for her comments and suggestions.

Conflits d'intérêts
Aucun à déclarer

Conflicts of Interest
None to declare

Édition/Editors: Erica Monteferrante & Aliya Affdal

Les éditeurs suivent les recommandations et les procédures décrites dans le Code of Conduct and Best Practice Guidelines for Journal Editors de COPE. Plus précisément, ils travaillent pour s'assurer des plus hautes normes éthiques de la publication, y compris l'identification et la gestion des conflits d'intérêts (pour les éditeurs et pour les auteurs), la juste évaluation des manuscrits et la publication de manuscrits qui répondent aux normes d'excellence de la revue.

The editors follow the recommendations and procedures outlined in the COPE Code of Conduct and Best Practice Guidelines for Journal Editors. Specifically, the editors will work to ensure the highest ethical standards of publication, including: the identification and management of conflicts of interest (for editors and for authors), the fair evaluation of manuscripts, and the publication of manuscripts that meet the journal's standards of excellence.

Évaluation/Peer-Review: Carl Mörch

Les recommandations des évaluateurs externes sont prises en considération de façon sérieuse par les éditeurs et les auteurs dans la préparation des manuscrits pour publication. Toutefois, être nommé comme évaluateurs n'indique pas nécessairement l'approbation de ce manuscrit. Les éditeurs de la *Revue canadienne de bioéthique* assument la responsabilité entière de l'acceptation finale et de la publication d'un article.

Reviewer evaluations are given serious consideration by the editors and authors in the preparation of manuscripts for publication. Nonetheless, being named as a reviewer does not necessarily denote approval of a manuscript; the editors of *Canadian Journal of Bioethics* take full responsibility for final acceptance and publication of an article.

REFERENCES

1. Miller D, Brown E. Artificial Intelligence in medical practice: the question to the answer? The American Journal of Medicine. 2018;131(2):129-133.
2. Nuffield Council on Bioethics. Artificial Intelligence (AI) in healthcare and research. Nuffield Council on Bioethics; 2018.
3. Challen R, Denny J, Pitt M, et al. Artificial intelligence, bias and clinical safety. BMJ Quality & Safety. 2019;28(3):231-237.
4. Hague D. Benefits, pitfalls, and potential bias in health care AI. North Carolina Medical Journal. 2019;80(4):219-223.
5. Akhtar A. New York is investigating UnitedHealth's use of a medical algorithm that steered black patients away from getting higher-quality care. Business Insider; 28 Oct 2019.
6. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366(6464):447-453.
7. Chen Y, Szolovits P, Ghassemi M. Can AI help reduce disparities in general medical and mental health care? AMA Journal of Ethics. 2019;21(2):E167-179.
8. Nordling L. A fairer way forward for AI in health care. Nature. 2019;573(7775):S103-S105.
9. van Ryn M, Burke J. The effect of patient race and socio-economic status on physicians' perceptions of patients. Social Science & Medicine. 2000;50(6):813-828.
10. Heinrich J, Nachum O. Identifying and correcting label bias in machine learning. arXiv. 2019;arXiv:1901.04966.
11. Howard A, Borenstein J. The ugly truth about ourselves and our robot creations: the problem of bias and social inequity. Science and Engineering Ethics. 2017;24(5):1521-1536.
12. Mittelstadt, B. Principles alone cannot guarantee ethical AI. Nature Machine Intelligence. 2019;1:501-507.
13. FitzGerald C, Hurst S. Implicit bias in healthcare professionals: a systematic review. BMC Medical Ethics. 2017;18:19.

14. Chapman E, Kaatz A, Carnes M. Physicians and implicit bias: how doctors may unwittingly perpetuate health care disparities. Journal of General Internal Medicine. 2013;28(11):1504-1510.
15. Wylie L, McConkey S. Insiders' insight: discrimination against Indigenous peoples through the eyes of health care professionals. Journal of Racial and Ethnic Health Disparities. 2019;6:37-45.
16. Reilly, J. Ogdie, A. et. al. Teaching about how doctors think: a longitudinal curriculum in cognitive bias and diagnostic error for residents. BMJ Quality & Safety 2013;22:1044-1050.
17. Gonzalez, C. Kim, M., Marantz, P. Implicit bias and its relation to health disparities: a teaching program and survey of medical students. Teaching and Learning in Medicine 2014;26(1):64-71.
18. Frey C, Osborne M. The future of employment. The Oxford Martin Programme on Technology and Employment. Working Paper. 2013.
19. Algorithmic Justice League. 2019.
20. The Montreal Declaration for the Responsible Development of Artificial Intelligence. Inven_T, University of Montreal; 2017.
21. The High-Level Expert Group on AI Guidelines. European Commission; 2019.
22. Price WN. Regulating black box medicine. Michigan Law Review. 2017;116(3):421-474.