BMJ Health & Care Informatics

A proposal for developing a platform that evaluates algorithmic equity and accuracy

Paul Cerrato,[1] John Halamka,[1] Michael Pencina[2]

To cite: Cerrato P, Halamka J, Pencina M. A proposal for developing a platform that evaluates algorithmic equity and accuracy. *BMJ Health Care Inform* 2022;**29**:e100423. doi:10.1136/bmjhci-2021-100423

Received 31 May 2021
Accepted 06 January 2022

[1]Paul Cerrato is Senior Research Analyst/Communications Specialist, Mayo Clinic Platform; John Halamka is President of Mayo Clinic Platform, Mayo Clinic Rochester, Rochester, Minnesota, USA
[2]Vice Dean for Data Science and Information Technology, Duke University, Durham, North Carolina, USA

Correspondence to
Paul Cerrato;
cerrato.paul@mayo.edu

ABSTRACT

We are at a pivotal moment in the development of healthcare artificial intelligence (AI), a point at which enthusiasm for machine learning has not caught up with the scientific evidence to support the equity and accuracy of diagnostic and therapeutic algorithms. This proposal examines algorithmic biases, including those related to race, gender and socioeconomic status, and accuracy, including the paucity of prospective studies and lack of multisite validation. We then suggest solutions to these problems. We describe the Mayo Clinic, Duke University, Change Healthcare project that is evaluating 35.1 billion healthcare records for bias. And we propose 'Ingredients' style labels and an AI evaluation/testing system to help clinicians judge the merits of products and services that include algorithms. Said testing would include input data sources and types, dataset population composition, algorithm validation techniques, bias assessment evaluation and performance metrics.

There have always been pivotal moments in the history of technology during which the enthusiasm for a specific innovation outpaces our ability to dispassionately evaluate its strengths and weaknesses. We are at that moment in the history of machine learning and its application in patient care. As clinicians and healthcare executives attempt to determine the role of machine learning-enhanced algorithms in the diagnosis, treatment, and prognosis of disease, many have raised this concern, questioning both the equity and accuracy of these sophisticated digital tools.

These concerns are now finding a voice in several recent guidelines. The Standard Protocol Items: Recommendations for Interventional Trials-Artificial Intelligence extension, a set of guidelines designed to help researchers develop AI-related clinical trials, states: 'It has been recognised that most recent AI studies are inadequately reported and existing reporting guidelines do not fully cover potential sources of bias specific to AI systems'.[1] Similarly, the Consolidated Standards of Reporting Trials-Artificial Intelligence extension, which serves as a guideline for reporting AI-related clinical trials explains: 'It has been shown that AI systems may be systematically biased towards different outputs, which may lead to different or even unfair treatment, on the basis of extant features'.[2]

HOW EXTENSIVE IS ALGORITHMIC BIAS?

There are numerous examples in healthcare that warrant the establishment of these guidelines. They fall into several distinct categories, including bias related to race, ethnic group, gender, socioeconomic status and geographic location; these inequities are impacting millions of lives. Obermeyer *et al*[3] have analysed a large, commercially available dataset used to determine which patients have complex health needs and require priority attention. In conjunction with a large academic hospital, the investigators identified 43 539 white and 6059 black primary care patients who were part of risk-based contracts. The analysis revealed that at any given risk score, blacks were considerably sicker than white patients, based on signs and symptoms. However, the commercial dataset did not recognise the greater disease burden in blacks because it was designed to assign risk scores based on total healthcare costs accrued in 1 year. Using this metric as a proxy for their medical need was flawed because the lower cost among blacks may have been due to less access to care, which in turn resulted from their distrust of the healthcare system and direct racial discrimination from providers.[4]

Gender bias has been documented in medical imaging datasets that have been used to train and test AI systems used for computer-assisted diagnosis. Larrazabal *et al*[5] studied the performance of deep neural networks used to diagnose 14 thoracic diseases using X-rays. When they compared gender-imbalanced datasets with datasets in which male and female candidates were equally represented,

they found that 'with a 25%/75% imbalance ratio, the average performance across all diseases in the minority class is significantly lower than a model trained with a perfectly balanced dataset'. Their analysis concluded that datasets that under-represent one gender results in biased classifiers, which in turn may lead to misclassification of pathology in the minority group. Their analysis is consistent with studies that have found women are less likely to receive high-quality care and more likely to die if they received suboptimal care.[6]

Similarly, there is evidence to suggest that machine learning enhanced algorithms that rely on electronic health record data under-represent patients in lower socioeconomic groups.[7] Typically, poorer patients receive fewer medications for chronic conditions and diagnostic tests and usually have less access to healthcare. This bias is likely to distort the advice being offered by clinical decision support systems that depend on these algorithms because said algorithms might give the impression that a specific disorder is uncommon in this patient subgroup, or that early interventions are unwarranted.

The inequities detected in healthcare-related algorithms mirror the biases observed in general purpose algorithms. One of the most well-known examples of these biases has been documented in an analysis of an online recruitment tool once used by the online retailer Amazon.[8] The algorithm was based on resumes that the retailer has collected over a decade and consisted primarily of white male candidates. In analysing this dataset, the digital tool was trained to look at word patterns in the resumes instead of relevant skill sets. As Lee *et al* explain: '…[T]hese data were benchmarked against the company's predominantly male engineering department to determine an applicant's fit. As a result, the AI software penalized any resume that contained the word "women's" in the text and downgraded the resumes of women who attended women's colleges, resulting in gender bias'. Similarly, there is evidence to demonstrate the existence of bias in online ads and facial recognition software, the latter having difficulty recognising darker-skinned complexions.

Of course, even a dataset that fairly represents all members of a targeted patient population is not very useful if it is inaccurate in other respects. A dataset that includes a representative sample of African-Americans, for instance, will be of limited value if the algorithm derived from that dataset is not validated with a second, external dataset. For example, when a machine learning approach was used to evaluate risk factors for *Clostridium difficile* infection, testing the algorithms in two different institutions found that the top 10 risk factors and top 10 protective factors were quite different between hospitals.[9]

Likewise, an algorithm that takes into account socioeconomic status may fall short if it is derived solely from retrospective analysis based on data that is not representative of the population to whom it will be applied. For example, randomised controlled trials (RCTs), which are the gold standard on which to base decisions about the effectiveness of any intervention, often do not enrol fully representative populations due to numerous inclusion and exclusion criteria. Carefully designed and well-executed analyses of 'real-world' datasets can supplement and expand the insights that can be derived from RCT data, especially in the creation of clinical decision support tools. The expectation that an algorithm will perform well on a local health system level today, requires evaluation of performance that incorporates the diversity of the current local population.

This highlights the importance of differentiating between algorithms that are supported by retrospective versus prospective research. There are hundreds of retrospective AI studies that have been mislabeled clinical trials, but in a recent review of the literature, we found only five RCTs that examined the value of machine learning and AI in patient care, and nine non-RCT prospective studies.[10] In light of these shortcoming, many healthcare providers hoping to implement algorithms with substantive evidence often turn to the US Food and Drug Administration (FDA) for guidance, working on the assumption that AI-enhanced software that has received FDA approval are more trustworthy and clinically proven to be safe and effective in patient care. Analysis of 130 FDA-approved AI devices suggests that the agency may not be able to perform an evaluation that guarantees the granularity that might be sought by local users.[11] Wu *et al* have found:

- Of the 130 FDA-approved AI devices, 126 relied solely on retrospective studies.
- Among the 54 high-risk devices evaluated, none included prospective studies.
- Of the 130 approved products, 93 did not report multisite evaluation.
- Fifty-nine of the approved AI devices included no mention of the sample size of the test population.
- Only 17 of the approved devices discussed a demographic subgroup.

This summary of recent FDA approvals demonstrates a significant limitation in the way AI-enhanced algorithms and devices are being evaluated. In addition, research projects that support a specific ML-enhanced algorithm also need to demonstrate that an algorithm's predictions are repeatable and reproducible. Similarly, the reference standard that is being used as 'ground truth' to evaluate an algorithm also has to be evidence-based. If, for example, a model compares a convolutional neural network's ability to identify diabetic retinopathy with the diagnostic skills of human ophthalmologists, there must be consensus from expert specialists on how to define diabetic retinopathy based on imaging data.

Pencina *et al* have enumerated several simple principles that need to be followed when constructing an algorithm-based clinical decision support tool.[12] It starts with the need to align target population to whom the model will be applied and the sample used to develop the model. For instance, the equations used to create the current national cholesterol guidelines are derived from persons who do not have the

Box 1 The fictitious product description could serve as a template for an artificial intelligence (AI) evaluation service that helps clinicians and healthcare executives make a more informed decision about how to invest in digital services that are equitable and accurate. The sample only includes a *few* of the most important algorithm features that can be documented in a 'nutrition label' style format. For clinicians with no background in information technology, an educational training session may be required to enable them to make useful comparisons among competing products. The graphic is a simplified version of what a product card might look like. It is intended to serve as the starting point for an iterative design process

RadiologyIntel
Summary: machine learning-based decision support software to augment medical imaging-related diagnosis of abdominal CT scans.
Data:
Input data sources: radiology information system/picture archiving and communication system, and epic electronic health record (EHR) system.
Input data type: digital abdominal images, text reports from radiologists, EHR narrative data on signs and symptoms, laboratory test results.
Training data location and time period: Acme Medical Center, Jamestown, Virginia, September 2014 to December 2016.
Statistical tests and metrics employed during training and validation testing
High-level Python-based neural network, Keras, TensorFlow.
Conducted on NVIDIA GeForce Graphical processing units.
Population composition
Ethnic composition
Non-Hispanic white 60%
Hispanic and Latino 18%
Black or African-American 13%
Asian 6%
Other 3%
Gender balance 55/45%, male/female
Primary outcome(s) XXX
Time horizon XXX
Algorithm and performance:
Type of algorithm employed
Convolutional neural network
Algorithm validation
Retrospective analysis*
Prospective clinical trial†
Size/Composition of training dataset:
55 000 inpatients at academic medical centre
Size/Composition of cross-validation dataset:
35 000 inpatients at community hospital
Performance metrics
Area under the curve 0.85
Sensitivity
Specificity
Classification accuracy 75%
Summary receiver operating curve 0.75
Bias assessment evaluation
Google TCAV
Audit-AI
Food and Drug Administration approval status

Continued

Box 1 Continued

510(k) Premarket approval—Approved December 2020
Warnings
This model is not intended to generate independent diagnostic decisions but is to be used as an adjunct to radiologist and attending physician's clinical expertise. Use of the algorithm should be discontinued if there are significant shifts in performance statistics or changes in patient population.
Published evidential support (fictitious references to illustrate the nutrition label model)
*Loretz A *et al*. Evaluation of an AI-based detection software in abdominal computed tomography scans. *JAMA* 2017;450:345–357.
†Mendez J *et al*. Randomised clinical trial to compare radiological imaging algorithm to radiologists' diagnostic skills. *Lancet* 2019;333:450–460.

disease, are between 40 and 79 years of age and are not taking lipid-lowering medication.[13] Using such a dataset to create algorithms that predict the likelihood of developing atherosclerotic cardiovascular disease among patients taking statins or who fall outside the age frame will incorrectly label many individuals as high and low risk. Likewise, careful selection and definition of the outcome of interest that aligns with the goals of care as well as one's choice of predictors to measure can influence the value of an algorithm to identify at-risk individuals. Furthermore, Pencina *et al* argue that given similar performance, preference should be given to simpler and more easily interpretable models. Finally, thorough evaluation of model performance consistent with the way the algorithm will be applied in practice is necessary.

Another problem that can generate biased predictions is putting too much emphasis on the 'average' patient and neglecting investigation of subgroup effects. Clinical studies need to perform the necessary subgroup analyses to detect the ethnic, gender or physiological characteristics of unrepresented groups that will then inform the development of clinical decision support algorithms. Several clinical trial re-analyses have documented these shortcomings, which we have summarised in an earlier publication.[14]

Finally, while it is important to take into account subgroup analyses when evaluating an AI-based algorithm, it is also important to emphasise that the accurate performance of an ML model within specific subgroups does not guarantee equity in the accrual of benefit. The evaluation must encompass the interplay of the model's output with the prevailing intervention allocation policy. Often, equity can be reached by adjusting the policy without diving too deeply into the algorithmic fairness of the model.

SOLUTIONS TO IMPROVE ALGORITHM TRANSPARENCY AND PERFORMANCE AND PROMOTE HEALTH EQUITY

Starting from the premise that any complex societal problem must first be measured before it can be solved, Mayo Clinic and Duke School of Medicine entered a collaboration with Optum/Change Healthcare focused on analysis of their data consisting of >35.1 billion

healthcare events and over 15.7 billion insurance claims to look for patterns of care and any possible inequities in that care. Change Healthcare provides social determinants of health, including economic vulnerability, education levels/gaps, race/ethnicity and household characteristics on about 125 million unique de-identified individuals. This provides a unique combined clinical and non-clinical view of healthcare journeys in the USA. A better understanding of this dataset will enable Mayo and Duke to design initiatives to help eradicate racism and offer services to underserved communities. One component of the project reviews the billing data, including ICD codes and CPT codes. It analyses diabetes care, as reflected by haemoglobin A1c testing and the use of telemedicine services, as well as planned study of the utilisation of colorectal cancer screening services, as reflected in the use of Cologuard, an at-home stool-DNA screening test (Mayo Clinic has a financial interest in Cologuard), colonoscopy and other screening methods. Utilisation of these services is being mapped against numerous social determinants of health when available, including a patient's education level, country of origin, economic stability indicator (financial), how likely they were to search for medical information on the internet, requests to their physician for information about medications, the presence of a senior adult in the household, number of children and home and car ownership.

The results of such analyses will help clinicians and healthcare executives develop more equitable digital tools, but they do not obviate the need to formally evaluate AI-enhanced algorithms and digital services to ensure that they achieve their stated purpose and help improve health equity. Unfortunately, the current digital solutions marketplace remains a 'Wild West' that is acutely in need of certifying protocols to address the aforementioned shortcomings. There are three possible pathways to follow in creating these evaluation services. One approach is to develop a system similar to the nutrition or drug label currently in place for most US foods and beverages and medications.[15] It would list many of the 'ingredients' that have been used to generate each algorithm or digital service, including how the dataset was derived and tested and what kind of clinical studies were conducted to demonstrate that it has value in routine patient care. It would also list the type of methodology used to develop the model, for example, convolutional neural network, random forest analysis, gradient boosting, the types of statistical tests and performance metrics that were used on the training and test sets and bias assessment tools employed. A second approach would be a *Consumer Reports*-like system. It would take a closer look at commercially available AI-enhanced services, outlining and comparing them much the way *Consumer Reports* compares appliances, automobiles and the like. This second approach would be facilitated by an across-health systems data and algorithm platform or federation where internal and external models can be tested, improved and selected. That would allow potential users to separate the wheat from the chaff, providing them with a reliable resource as they decide how to make investments. A third approach would be a hybrid evaluation system that combined elements of the first two systems.

Applying these types of evaluation tools to existing diagnostic and screening algorithms might avert the poor model performances that have been reported in the medical literature. For example, an analysis of the Epic Deterioration Index, which was designed to identify subgroups of hospitalised patients with COVID-19 at risk for complications and alert clinicians to the onset of sepsis, fell short of expectations.[16] The system had to be deactivated 'because of spurious alerting owing to changes in patients' demographic characteristics associated with the COVID-19 pandemic'.[17]

For any of these approaches to be successful, it is necessary to develop an AI evaluation system with specific evaluation criteria and testing environments to judge model performance and impact on health equity. The best place to start is by taking a critical look at the input data being collected for each dataset. Any algorithm developer interested in demonstrating that they have a representative service will want to present statistics on the percentages of white, black, Asian, Hispanic and other groups in their dataset, as illustrated in box 1 and table 1. Similarly, they will attest to its male/female balance, as well as its socioeconomic and geographic breakdowns. It is also important to keep in mind that an equitable algorithm must be derived from a dataset that is representative of the entire population to be served. The AI evaluation system described here would create standards by which a product can be evaluated. There would then be multiple testing labs available, as well as several certification entities that use the results of these labs.

This form of algorithmic hygiene is a bare minimum standard, however. There are numerous types of bias that require attention, including statistical overestimation and underestimation, confirmation bias and anchoring bias. In addition, developers also need to be realistic about how data are entered into their training set. Electronic and human data entry can inadvertently insert biased information into a dataset's raw data. Many types of healthcare require humans to enter descriptors and tags that may be influenced by their own prejudices and stereotypes. And even devices like rulers, cameras and voice recognition software used to generate data can enter biased data. Alegion, a company that does ground truth training for machine learning initiative, points out 'For example, a camera with a chromatic filter will generate images with a consistent colour bias. An 11-7/8 inch long "foot ruler" will always over-represent lengths'.[18]

Vendors will also want to take the next step and demonstrate that the composition of their data scientist team is diverse and represents all the segments of society that have often been under-represented in healthcare. Without such a diverse team, subtle choices made during the data collection process will produce unbalanced datasets. Additional credentialling documents that will allow

Table 1 Clinical AI reports

Name of device or algorithm	Brief description	Data collection methods	FDA approval status	Type of algorithm	Data set composition	Population ethnic composition	Bias assessment evaluation	Model evaluation/ Research protocol	Metrics for performance errors* †	Clinical workflow implementation
RadiologyIntel	Decision support software to augment medical imaging-related diagnosis	Standard H&E stained images, stimulated Raman histology	510(k) Premarket notification	Convolutional neural network	Size/ Composition of training dataset: 550 000 inpatients, academic medical centres Size/ Composition of testing dataset: 350 000 inpatients at community hospitals	Non-Hispanic white 60% Hispanic and Latino 18% Black/African-American 13% Asian 6% Other 3%	Google TCAV Audit-AI	Multi-centred prospective clinical trial and retrospective analysis	Area under the curve 0.85 Classification accuracy 75%	Integrated into 50 hospitals via EHR systems, including Epic, Cerner
DiabetEYE	CDS system to enhance screening/ diagnosis of diabetic retinopathy	Widefield stereoscopic photography and macular optical coherence tomography	De novo pathway	Convolutional neural network	Size/ Composition of training dataset: 7000 outpatients, primary care clinic Size/ Composition of testing dataset: 5000 Outpatients at independent clinic	Non-Hispanic white 70% Hispanic and Latino 10% Black/African-American 10% Other 10%	None available	Randomised controlled trial	Sensitivity, 81%, specificity, 90%, Area under the curve 0.80 Confusion matrix 0.91	Implemented in 150 primary care clinics in the USA

*Mishra.[19]
†Scott *et al*.[20]
AI, artificial intelligence.

Box 2 Bias detection analytics tools

Although it is virtually impossible to eliminate all bias from artificial intelligence (AI)-based datasets and algorithms, there are several tools that can help mitigate the problem. These tools are essentially algorithmic solutions to correct algorithmic inequities. Here are a few examples of these detection tools.

Testing with concept activation vectors

Testing with concept activation vectors (TCAV) is one of Google's tools to address algorithmic bias, including bias by race, gender and location. For example, in a neural network-based system designed to classify images and identify a zebra, TCAVs can determine how sensitive the presence of stripes are in predicting the presence of the animal.[21] The tool uses directional derivatives to estimate the degree to which a user-defined concept is important to the results of the classification task at hand. Using concept activation vectors can help detect biases by unearthing unexpected word, class, of concept associations that suggest an inequity. In one analysis, for instance, the 'female' concept was linked to the 'apron' class.[22]

Audit-AI

Makes use of a Python library from pymetrics that can detect discrimination by locating specific patterns in the training data. For example, it can input mammography access data for various ethnic groups into an algorithm in question to generate proportional pass rates of various groups, comparing white with black patients. The resulting bias ratio can then be analysed statistically looking for significant differences and clinical meaningful differences in healthcare access.[23]

AI Fairness 360

A Python-based bias detection algorithm from IBM, AI Fairness 360 (AIF360) starts with the assumption that many datasets do not contain enough diverse data points. The IBM team explains 'Bias detection is demonstrated using several metrics, including disparate impact, average odds difference, statistical parity difference, equal opportunity difference and Theil index. Bias alleviation is explored via a variety of methods, including reweighing (preprocessing algorithm), prejudice remover (in-processing algorithm) and disparate impact remover (preprocessing technique)'. A use case of how AIF360 can be used to reveal discrimination is a scoring model that looks at healthcare utilisation.[24]

Tariq *et al* have also reviewed numerous AI evaluation tools that are worth considering.[25] They have developed a 10-question tool to evaluate AI products that include 'model type, dataset size and distribution, dataset demographics/subgroups, standalone model performance, comparative performance against a gold standard, failure analysis, publications, participation in public challenges, dataset release and scale of implementation'.

the best solutions providers to stand out would include bias impact statements, inclusive design principles, algorithm auditing process and cross-functional work teams. Algorithm developers can also use several analytical tools designed to detect such problems, including Google's TCAV, Audit-AI and IBM's AI Fairness 360, discussed in box 2.

The history of medicine is filled with 'near misses', technologies that had the potential to improve patient care but that failed to hit their intended target and did not live up to that potential once rigorously tested. The evidence suggests that machine learning-enhanced algorithms as a group do not fall into that category; instead, they are poised to profoundly transform the diagnosis, treatment and prognosis of disease. As we have documented in earlier publications,[10] there are a small number of RCTs and non-RCT prospective studies to support the use of these digital tools in several medical specialties, including oncology, radiology, ophthalmology and dermatology. But for clinicians and healthcare executives to make decisions regarding commercially available algorithmic services, we propose an evaluation platform that dispassionately reports on the basic features of each product. Such a platform would allow providers to compare competing products and choose those that are equitable and accurate.

Twitter Paul Cerrato @plcerrato

Contributors All authors were the contributors to the paper.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval This study does not involve human participants.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data sharing not applicable as no datasets generated and/or analysed for this study. Data sharing not applicable.

REFERENCES

1 Cruz Rivera S, Liu X, Chan A-W, *et al*. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Nat Med* 2020;26:1351–63.
2 Liu X, Cruz Rivera S, Moher D, *et al*. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat Med* 2020;26:1364–74.
3 Obermeyer Z, Powers B, Vogeli C, *et al*. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 2019;366:447–53.
4 Ledford H. Millions of black people affected by racial bias in healthcare algorithms. *Nature* 2019;574:608–9.
5 Larrazabal AJ, Nieto N, Peterson V, *et al*. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proc Natl Acad Sci U S A* 2020;117:12592–4.
6 Li S, Fonarow GC, Mukamal KJ, *et al*. Sex and Race/Ethnicity-Related disparities in care and outcomes after hospitalization for coronary artery disease among older adults. *Circ Cardiovasc Qual Outcomes* 2016;9:S36–44.
7 Gianfrancesco MA, Tamang S, Yazdany J, *et al*. Potential biases in machine learning algorithms using electronic health record data. *JAMA Intern Med* 2018;178:1544–7.
8 Lee NC, Resnick P, Barton G. Algorithmic bias detection and mitigation: best practices and policies to reduce consumer harms. Brookings institution, 2019. Available: https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/#footnote-8
9 Oh J, Makar M, Fusco C, *et al*. A generalizable, data-driven approach to predict daily risk of Clostridium difficile infection at two large academic health centers. *Infect Control Hosp Epidemiol* 2018;39:425–33.
10 Halamka J, Cerrato P. The digital reconstruction of health care. *NEJM Catalyst* 2020;1.
11 Wu E, Wu K, Daneshjou R, *et al*. How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. *Nat Med* 2021;27:582–4.
12 Pencina MJ, Goldstein BA, D'Agostino RB. Prediction Models - Development, Evaluation, and Clinical Application. *N Engl J Med* 2020;382:1583–6.

13 Goff DC, Lloyd-Jones DM, Bennett G. 2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American College of Cardiology/American heart association Task force on practice guidelines. *Circulation* 2014;129:S49–73.
14 Cerrato P, Halamka J. *Redefining clinical decision support: data analytics, artificial intelligence, and diagnostic reasoning*. Boca Raton, FL: Taylor & Francis/HIMSS, 2020.
15 Sendak M, Elish MC, Gao M. The Human Body is a Black Box": Supporting Clinical Decision-Making with Deep Learning. *arXiv* 2019:1911.08089.
16 Singh K, Valley TS, Tang S, *et al*. Evaluating a widely implemented proprietary deterioration index model among hospitalized patients with COVID-19. *Ann Am Thorac Soc* 2021;18:1129–37.
17 Finlayson SG, Subbaswamy A, Singh K, *et al*. The clinician and dataset shift in artificial intelligence. *N Engl J Med* 2021;385:283–6.
18 Editorial team. 4 Sources of Machine Learning Bias & How to Mitigate the Impact on AI Systems. Inside Big Data, 2018. Available: https://insidebigdata.com/2018/08/20/machine-learning-bias-ai-systems/
19 Mishra A. Metrics to evaluate your machine learning algorithm. towards data science, 2018. Available: https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234
20 Scott I, Carter S, Coiera E. Clinician checklist for assessing suitability of machine learning applications in healthcare. *BMJ Health Care Inform* 2021;28:e100:e100251.
21 Asokan A. *Top 5 tools data scientists can use to mitigate biases in algorithms*. Analytics India Magazine, 2019. https://analyticsindiamag.com/top-5-tools-data-scientists-can-use-to-mitigate-biases-in-algorithms/
22 Kim B, Wattenberg M, Gilmer G. Interpretability beyond feature attribution: quantitative testing with concept activation vectors (TCAV). Proceedings of the 35th International Conference on machine learning, Stockholm, Sweden, PMLR 80, 2018. Available: http://proceedings.mlr.press/v80/kim18d/kim18d.pdf
23 Pymetrics/ audit AI., 2020. Available: https://github.com/pymetrics/audit-ai [Accessed 02 Apr 2021].
24 Varshney KR. Introducing AI fairness 360, 2018. Available: https://www.ibm.com/blogs/research/2018/09/ai-fairness-360/
25 Tariq A, Purkayastha S, Padmanaban G. Reading race: AI recognises patient's racial identity in medical images. *J Am Coll Radiol* 2020;17:1371–81.