Health Data in Practice lecture series

Trustworthy development and use of Artificial Intelligence in health care

Carol Dezateux

8th December 2020

Institute of Population Health Sciences

wellcome

Queen Mary University of London
Barts and The London

Learning Objectives

At the end of this lecture on trustworthy development and use of AI in health care you will be able to

- Understand emerging approaches to ensure equitable development and use of AI algorithms
- Summarise potential biases encountered in development and use of AI algorithms in health care
- Understand existing and planned modifications to Equator standards for clinical trials (SPIRIT-AI, CONSORT-AI), diagnostic tests (STARD-AI) and prediction models (TRIPOD-ML)
- Be able to apply these concepts to studies of the development and use of AI in health care
- Be able to access key documents and reports and maintain awareness of developments in relevant standards and codes of conduct

UK public sector failing to be open about its use of AI, review finds

Natasha Lomas @riptari / 2:56 PM GMT • February 10, 2020 Comment

Image Credits: Roy Scott / Getty Images

A report into the use of artificial intelligence by the U.K.'s public sector has warned that the government is failing to be open about automated decision-making technologies which have the potential to significantly impact citizens' lives.

Ministers have been especially bullish on injecting new technologies into the delivery of taxpayer-funded healthcare — with health minister Matt Hancock setting out a tech-fueled vision of "preventative, predictive and personalised care" in 2018, calling for a root and branch digital transformation of the National Health Service (NHS) to support piping patient data to a new generation of "healthtech" apps and services.

AI migration tools could have inhumane consequences

By E&T editorial staff

Published Wednesday, March 18, 2020

A University of Exeter study has suggested that AI could "revolutionise" how international migration is managed, but warned that it could also reinforce inhumane practices and permitting discrimination.

AI is being used more and more by governments and international organisations preparing for and managing mass migration, such as by performing identity checks at borders, processing data about visa and asylum applicants.

Search jobs My account Search UK edition
The Guardian

Google

More than 1,200 Google workers condemn firing of AI scientist Timnit Gebru

More than 1,500 researchers also sign letter after Black expert on ethics says Google tried to suppress her research on bias

Julia Carrie Wong *in San Francisco and agencies*

@juliacarriew Email

Fri 4 Dec 2020 19.48 GMT

2,390

Timnit Gebru in San Francisco in 2018. Photograph: Kimberly White/Getty Images for TechCrunch

More than 1,200 Google employees and more than 1,500 academic researchers are speaking out in protest after a prominent Black scientist studying the ethics of artificial intelligence said she was fired by Google after the company attempted to suppress her research and she criticized its diversity efforts.

Equitable development and use of AI

Published in final edited form as:
Ann Intern Med. 2018 December 18; 169(12): 866–872. doi:10.7326/M18-1990.

Ensuring Fairness in Machine Learning to Advance Health Equity

Alvin Rajkomar, MD[*],
Google, Mountain View, and University of California, San Francisco, San Francisco, California

Michaela Hardt, PhD[*],
Google, Mountain View, California

Michael D. Howell, MD, MPH,
Google, Mountain View, California

Greg Corrado, PhD, and
Google, Mountain View, California

Marshall H. Chin, MD, MPH
University of Chicago, Chicago, Illinois

Equitable development and use of AI: recommendations

Design

- Determine the goal of a machine-learning model and review it with diverse stakeholders, including protected groups.
- Ensure that the model is related to the desired patient outcome and can be integrated into clinical workflows.
- Discuss ethical concerns of how the model could be used.
- Decide what groups to classify as protected.
- Study whether the historical data are affected by health care disparities that could lead to label bias. If so, investigate alternative labels.

Data collection

- Collect and document training data to build a machine-learning model.
- Ensure that patients in the protected group can be identified (weighing cohort bias against privacy concerns).
- Assess whether the protected group is represented adequately in terms of numbers and features.

Ann Intern Med. 2018 December 18; 169(12): 866–872. doi:10.7326/M18-1990.

Equitable development and use of AI: recommendations

Training

- Train a model taking into account the fairness goals.

Evaluation

- Measure important metrics and allocation across groups.
- Compare deployment data with training data to ensure comparability.
- Assess the usefulness of predictions to clinicians initially without affecting patients.

Launch review

- Evaluate whether a model should be launched with all stakeholders, including representatives from the protected group.
- Monitored deployment
- Systematically monitor data and important metrics throughout deployment.
- Gradually launch and continuously evaluate metrics with automated alerts.
- Consider a formal clinical trial design to assess patient outcomes.
- Periodically collect feedback from clinicians and patients.

Biases in AI: model design

- **Label bias:**
  A label that does not mean the same thing for all patients because it is an imperfect proxy that is subject to health care disparities rather than an adjudicated truth. This is a generalization of test-referral and test-interpretation bias in the statistics literature

- **Cohort bias:**
  Defaulting to traditional or easily measured groups without considering other potentially protected groups or levels of granularity (e.g., whether sex is recorded as male, female, or other or more granular categories)

Ann Intern Med. 2018 December 18; 169(12): 866–872. doi:10.7326/M18-1990.

Biases in AI: training data

- **Minority bias**:
  The protected group may have insufficient numbers of patients for a model to learn the correct statistical patterns
- **Missing data bias**:
  Data may be missing for protected groups in a nonrandom fashion, which makes an accurate prediction hard to render (e.g., a model may underdetect clinical deterioration in patients under contact isolation because they have fewer vital signs)
- **Informativeness bias:**
  Features may be less informative to render a prediction in a protected group (e.g., identifying melanoma from an image of a patient with dark skin may be more difficult)
- **Training–serving skew:**
  The model may be deployed on patients whose data are not similar to the data on which the model was trained. The training data may not be representative (i.e., selection bias), or the deployment data may differ from the training data (e.g., a lack of unified methods for data collection or not recording data with standardized schemas)

Ann Intern Med. 2018 December 18; 169(12): 866–872. doi:10.7326/M18-1990.

Biases in AI: interactions with clinicians

- **Automation bias:**
  If clinicians are unaware that a model is less accurate for a specific group, they may trust it too much and inappropriately act on inaccurate predictions
- **Feedback loops:**
  If the clinician accepts the recommendation of a model even when it is incorrect to do so, the model's recommended versus administered treatments will always match. The next time the model is trained, it will learn to continue these mistakes
- **Dismissal bias:**
  Conscious or unconscious desensitization to alerts that are systematically incorrect for a protected group (e.g., an early-warning score for patients with sepsis). Alert fatigue is a form of this
- **Allocation discrepancy:**
  If the protected group has disproportionately fewer positive predictions, then resources allocated by the predictions (e.g., extra clinical attention or social services) are withheld from that group

Ann Intern Med. 2018 December 18; 169(12): 866–872. doi:10.7326/M18-1990.

Biases in AI: interactions with patients

- **Privilege bias:**
  Models may be unavailable in settings where protected groups receive care or require technology/sensors disproportionately available to the nonprotected class

- **Informed mistrust:**
  Given historical exploitation and unethical practices, protected groups may believe that a model is biased against them. These patients may avoid seeking care from clinicians or systems that use the model or deliberately omit information. The protected group may be harmed by not receiving appropriate care

- **Agency bias:**
  Protected groups may not have input into the development, use, and evaluation of models. They may not have the resources, education, or political influence to detect biases, protest, and force correction

Ann Intern Med. 2018 December 18; 169(12): 866–872. doi:10.7326/M18-1990.

Distributive justice options for machine learning / AI

- **Equal patient outcomes:** The model should lead to equal patient outcomes across groups
- **Equal performance:** The model performs equally well across groups for such metrics as accuracy, sensitivity, specificity, and positive predictive value
- **Equal allocation:** Allocation of resources as decided by the model is equal across groups, possibly after controlling for all relevant factors

Ann Intern Med. 2018 December 18; 169(12): 866–872. doi:10.7326/M18-1990.

UK Government code of conduct for data-driven health and care technology

Principle 1: Understand users, their needs and the context

Principle 2: Define the outcome and how the technology will contribute to it

Principle 3: Use data that is in line with appropriate guidelines for the purpose for which it is being used

Principle 4: Be fair, transparent and accountable about what data is being used

Principle 5: Make use of open standards

Principle 6: Be transparent about the limitations of the data used

Principle 7: Show what type of algorithm is being developed or deployed, the ethical examination of how the data is used, how its performance will be validated and how it will be integrated into health and care provision

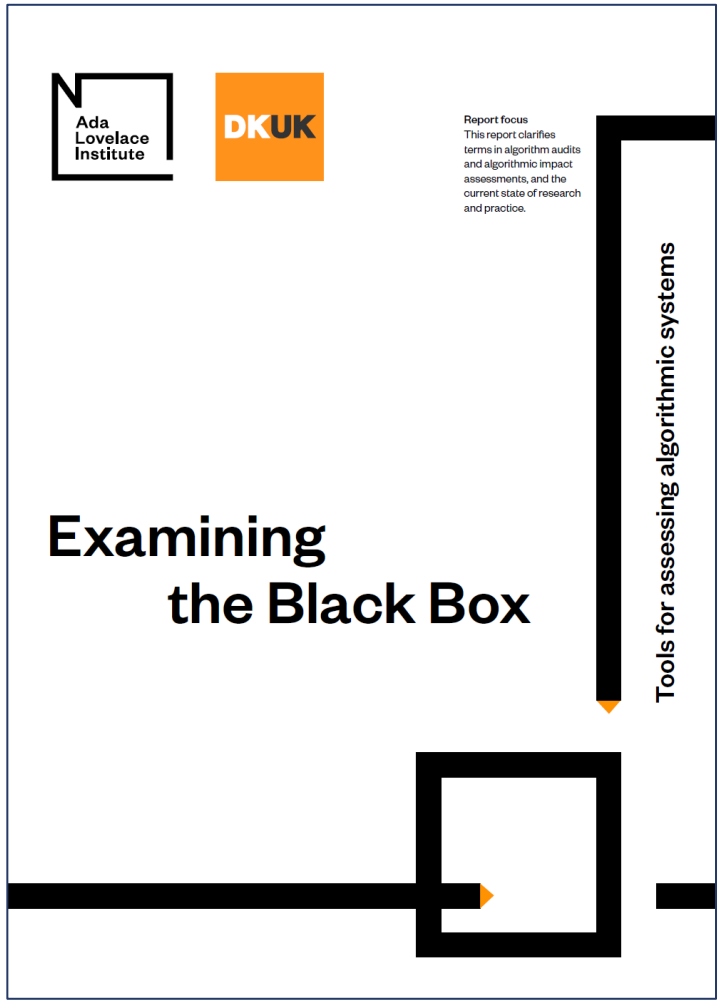
Principle 8: Generate evidence of effectiveness for the intended use and value for money

Principle 9: Make security integral to the design

Principle 10: Define the commercial strategy

Source: https://www.gov.uk/government/publications/code-of-conduct-for-data-driven-health-and-care-technology/initial-code-of-conduct-for-data-driven-health-and-care-technology

Two terms, four approaches for assessing algorithmic systems



	Algorithm audits		Algorithmic impact assessments	
	Bias Audit	**Regulatory inspection**	**Algorithmic risk assessment**	**Algorithmic impact evaluation**
What?	A targeted approach focused on assessing algorithmic systems for bias	A broad approach focussed on an algorithmic system's compliance with regulation or norms, and requiring a number of different tools and methods	Assessing possible societal impacts of an algorithmic system before the system is in use (with ongoing monitoring advised)	Assessing possible societal impacts of an algorithmic system on the users or population it affects after it is in use
When?	After deployment	After deployment, potentially ongoing	Before deployment, potentially ongoing	After deployment
Who by?	Researchers, investigative journalists, data scientists	Regulators, auditing and compliance professionals	Creators or commissioners of the algorithmic system	Researchers, policymakers
Origin	Social science audit studies	Regulatory auditing in other fields e.g. financial audits	Environmental impact assessments, data protection impact assessments	Policy impact assessments, which typically are evaluative after the fact
Case study	'Gender shades' study of bias in classification by facial recognition APIs	UK Information Commissioner's Office AI auditing framework draft guidance	Canadian Government's algorithmic impact assessment	Stanford's 'Impact evaluation of a predictive risk modeling tool for Allegheny County's Child Welfare Office'
Status	More established methodology in algorithm context; limited scope	Emerging methodology, skills and capacity requirements for regulators, more established approaches for compliance teams in tech sector	Some established methodologies in other fields, new to algorithm context; requiring evidence as to its applicability and best practice	Established methodology new to algorithm context; requiring evidence as to its applicability and best practice

https://www.adalovelaceinstitute.org/report/examining-the-black-box-tools-for-assessing-algorithmic-systems/

Two terms, four approaches for assessing algorithmic systems

Timing

Before and during deployment
- Algorithmic risk assessment

After deployment
- Bias audit
- Regulatory inspection
- Algorithmic impact evaluation

	Algorithm audits		Algorithmic impact assessments	
	Bias Audit	**Regulatory inspection**	**Algorithmic risk assessment**	**Algorithmic impact evaluation**
What?	A targeted approach focused on assessing algorithmic systems for bias	A broad approach focussed on an algorithmic system's compliance with regulation or norms, and requiring a number of different tools and methods	Assessing possible societal impacts of an algorithmic system before the system is in use (with ongoing monitoring advised)	Assessing possible societal impacts of an algorithmic system on the users or population it affects after it is in use
When?	After deployment	After deployment, potentially ongoing	Before deployment, potentially ongoing	After deployment

https://www.adalovelaceinstitute.org/report/examining-the-black-box-tools-for-assessing-algorithmic-systems/

RESEARCH METHODS AND REPORTING

Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI Extension

Xiaoxuan Liu,[1,2,3,4,5] Samantha Cruz Rivera,[5,6] David Moher,[7,8] Melanie J Calvert,[4,5,6,9,10,11] Alastair K Denniston,[1,2,4,5,6,12] On behalf of the SPIRIT-AI and CONSORT-AI Working Group

For numbered affiliations see end of the article.

Correspondence to:
A K Denniston, Institute of Inflammation and Ageing, College of Medical and Dental Sciences, University of Birmingham, Birmingham, UK,
a.denniston@bham.ac.uk

Cite this as: *BMJ* 2020;370:m3164
http://dx.doi.org/10.1136/bmj.m3164

Accepted: 4 August 2020

The CONSORT 2010 (Consolidated Standards of Reporting Trials) statement provides minimum guidelines for reporting randomised trials. Its widespread use has been instrumental in ensuring transparency when evaluating new interventions. More recently, there has been a growing recognition that interventions

intervention, including instructions and skills required for use, the setting in which the AI intervention is integrated, the handling of inputs and outputs of the AI intervention, the human-AI interaction and providing analysis of error cases.

CONSORT-AI will help promote

correspondence

Developing specific reporting guidelines for diagnostic accuracy studies assessing AI interventions: The STARD-AI Steering Group

To the Editor — Artificial intelligence (AI)-based technologies dominate medical headlines and are routinely touted as the panacea for a number of longstanding deficiencies across health systems globally. Stakeholders from healthcare, government, computer science and industry backgrounds

interpretation (e.g., the use of external validation datasets, complexities of datasets and comparison to human performance) and the lack of standardized nomenclature (e.g., the definition of a 'validation dataset'), as well as the heterogeneity of outcome measures (e.g., area under the receiver

Furthermore, as a central aspect of our guideline development, we have engaged groups from typically underrepresented regions, such as Asia and Africa, in order to ensure that the AI extension to STARD will be viewed as applicable across a global scale.

RESEARCH METHODS AND REPORTING

Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI Extension

Samantha Cruz Rivera,[1,2] Xiaoxuan Liu,[2,3,4,5,6] An-Wen Chan,[7] Alastair K Denniston,[1,2,3,4,5,8] Melanie J Calvert,[1,2,6,9,10,11] On behalf of the SPIRIT-AI and CONSORT-AI Working Group

For numbered affiliations see end of the article.

Correspondence to:
A K Denniston,
Institute of Inflammation and Ageing, College of Medical and Dental Sciences, University of Birmingham, Birmingham, UK,
a.denniston@bham.ac.uk

Cite this as: *BMJ* 2020;370:m3210
http://dx.doi.org/10.1136/bmj.m3210

Accepted: 4 August 2020

The SPIRIT 2013 (The Standard Protocol Items: Recommendations for Interventional Trials) statement aims to improve the completeness of clinical trial protocol reporting, by providing evidence-based recommendations for the minimum set of items to be addressed. This guidance has been

investigators provide clear descriptions of the AI intervention, including instructions and skills required for use, the setting in which the AI intervention will be integrated, considerations around the handling of input and output data, the human-AI interaction and analysis of error cases.

Reporting of artificial intelligence prediction models

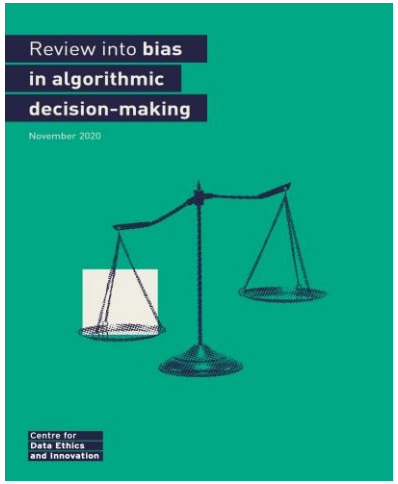
Data-driven technologies that form the basis of the digital health-care revolution provide potentially important opportunities to deliver improvements

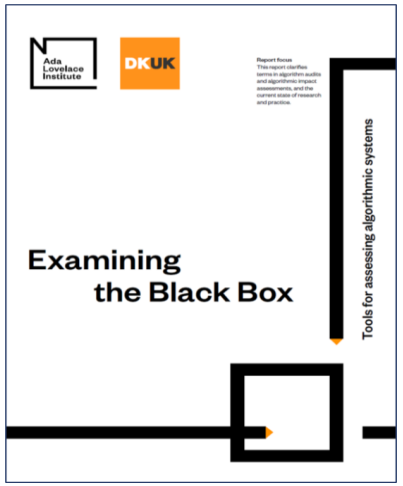
in individual care and to advance innovation in medical research. Digital health technologies include mobile devices and health apps (m-health), e-health

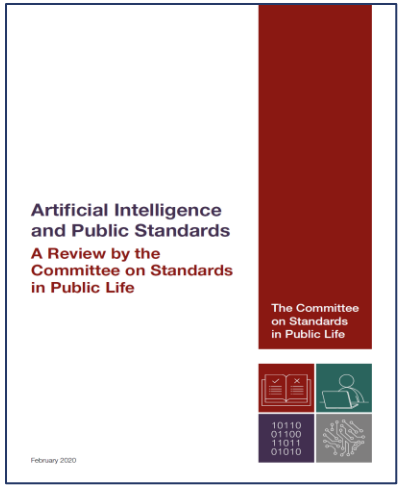
www.thelancet.com Vol 393 April 20, 2019

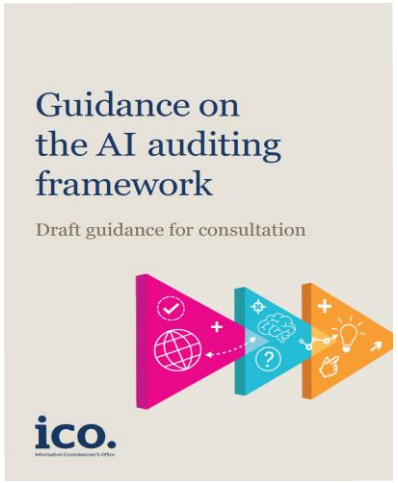
https://www.equator-network.org/reporting-guidelines/

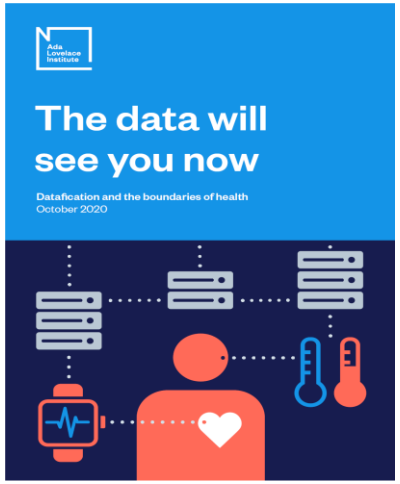
Further reading: key reports on AI development, use, and ethics

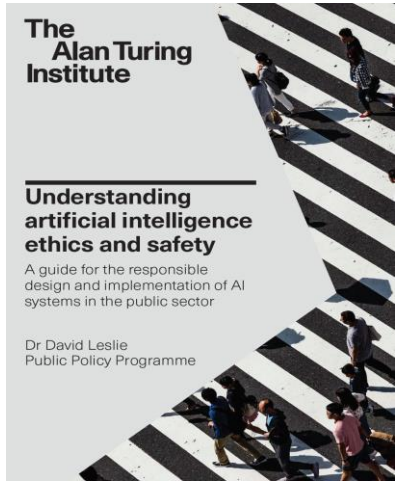












All on QM Plus as well as slides & published papers