BMJ Health & Care Informatics

'Improving smart medication management': an online expert discussion

David W Bates,[1] Hsiang-Yin Cheng,[2] NT Cheung,[3] Rita Jew,[4] Fraz Mir,[5] Robyn Tamblyn,[6] Yu-Chuan Li [7]

To cite: Bates DW, Cheng H-Y, Cheung NT, *et al*. 'Improving smart medication management': an online expert discussion. *BMJ Health Care Inform* 2022;**29**:e100540. doi:10.1136/bmjhci-2021-100540

Received 04 January 2022
Accepted 18 April 2022

[1]Division of General Internal Medicine and Primary Care, Brigham and Women's Hospital, Boston, Massachusetts, USA
[2]Taipei Medical University, Taipei, Taiwan
[3]Hong Kong Hospital Authority, Hong Kong, Hong Kong
[4]ISMP, Horsham, Pennsylvania, USA
[5]Addenbrooke's Hospital, Cambridge, UK
[6]McGill University, Montreal, Québec, Canada
[7]Taipei Medical University, Taipei, UK

Correspondence to
Professor David W Bates;
dbates@bwh.harvard.edu

ABSTRACT

Medication safety continues to be a problem inside and outside the hospital, partly because new smart technologies can cause new drug-related challenges to prescribers and patients. Better integrated digital and information technology (IT) systems, improved education on prescribing for prescribers and greater patient-centred care that empowers patients to take control of their medications are all vital to safer and more effective prescribing. In July 2021, a roundtable discussion was held as a spin-off meeting of the International Forum on Quality and Safety in Health Care Europe 2021 to discuss challenges and future direction in smart medication management. This manuscript summarises the discussion focusing on the aspects of digital and IT systems, safe prescribing, improved communication and education, and drug adherence.

INTRODUCTION

Medications are a cornerstone in patient management in primary, secondary and tertiary care. However, with about 9% of prescriptions containing errors[1] and patients often taking their prescribed medications incorrectly or not at all, medication safety continues to be a problem inside and outside the hospital.

As the baby boomer generation enters their senior years, the population that needs most medications is expected to double by 2036, when one in four persons will be 65 or older.[2] This trend is present in many industrialised countries, where both health and social policy efforts are being mobilised to reduce preventable morbidity that leads to healthcare use and loss of independence.[3 4]

In recent years, apps and other digital tools have been implemented in healthcare systems to assist in drug management. Yet, these new smart technologies can cause new challenges to prescribers, nurses, pharmacists and patients. Healthcare systems and staff need to ensure correct prescriptions and that patients take their medications as prescribed and report if side effects occur. To achieve improved outcomes for patients, human factors are as important as technology's role.

In July 2021 a roundtable discussion was held as a spin-off meeting of the International Forum on Quality and Safety in Health Care Europe 2021 to discuss challenges and future direction in smart medication management.

This manuscript summarises the discussion focusing on the aspects of digital and information technology (IT) systems, safe prescribing, improved communication and education, and drug adherence.

IMPROVING DIGITAL AND IT SYSTEMS FOR SAFER PRESCRIBING

In many developed countries, prescriptions are nowadays written mostly electronically. This allows them to be checked for safety related problems such as drugs that interact, allergies, doses that are too high or too low, and appropriateness of dosing in patients with conditions such as chronic renal insufficiency.

One of the main drivers for developing and implementing electronic medical records (EMRs) systems has been the promise of improved healthcare quality, using tools like Computerised Physician Order Entry and Clinical Decision Support (CDS).[5] The ability of EMRs, especially CDS, to improve medication safety has been demonstrated[6 7] and their transformative potential shown.[8]

However, more recently, it has been established that drug alerts as part of CDS being delivered routinely appear to result in almost no benefit. This has occurred with the almost complete conversion in the USA to commercial drug knowledge and alert applications. For example, one study[9] showed that the effectiveness of warnings about drug interactions fell dramatically after conversion to a commercial drug knowledge system. Another study[10] demonstrated that among about 5000 warnings about renal dosing, physicians responded to none of them. A third[11] showed that high-priority drug–drug interaction alerts were regularly overridden, probably because clinicians were getting so many warnings that they developed alert fatigue and ignored even the most important.

There can be many 'unintended consequences'[12] which may include increased risk of medication errors, or new types of errors.[13] Poorly designed or implemented EMRs are widely implicated in clinician burn-out[14] which can also lead to poorer quality of healthcare.[15]

How then can the original goal of improved healthcare quality and medication safety through EMRs be achieved? The aviation industry provides a good example. A sustained focus on safety throughout the industry has transformed the inherently unsafe activity of flying into one of the safest forms of transportation in the world. Although the flying machines themselves have improved, consideration of human factors in designing cockpits, the development of safety procedures and continuous monitoring of risks and incidents has allowed continuous improvements in safety over the decades.[16]

To achieve similar success in medication safety, two areas of improvement are important. First, usability and human factors are critical to building safe and effective medication management in EMRs with end user input in CDS design increasing the likelihood of it being successful.[17] Some features have been verified that increased the alert's perceived utility and can be used to improve effectiveness and reduce omitted warnings, for example, in a CDS tool targeting QT-interval prolonging medications.[18]

Second, a whole systems approach is needed. EMRs are complicated tools, being deployed in the complex healthcare delivery environment. The effects of interventions may be impossible to predict. Just focusing on the prescribing physician and the prescription will not be sufficient. All the processes, and all the healthcare providers involved in medication management in a patient's healthcare journey must be considered. In particular, the role of the patients themselves has been underexplored in the field of medication safety.

There are also other areas that represent ongoing challenges in safe prescribing. Medication lists are often incomplete, and there is still no clear approach to getting the most accurate medication list. Patients' role and common medication process practices agreed in hospital are crucial to ensure medication lists are up to date.

Difficulties also persist in writing prescriptions with complex descriptions of how the patient should take medications for example, prednisone that often requires stepwise tapering. In the inpatient and outpatient setting better approaches are needed to enable different specialties and all parts of the hospital team, including nursing, pharmacy, physicians respectively to be 'on the same page' regarding what the patient is taking. EMR records need to be implemented into practice and healthcare providers need to be educated how to use them in the medication process.

Still, the most obvious challenge—especially given the huge costs which have been expended on developing EMRs—is getting the point-of-care CDS to deliver important suggestions to clinicians—yet not bombard them with unimportant warnings. This issue represents a burning platform if EMRs are to realise their benefits on the medication safety front.

IMPROVING EDUCATION AND COMMUNICATION REGARDING PRESCRIBING

All prescribers must have a basic understanding of the medicines they prescribe to their patients. This traditionally includes knowing the indication for a particular drug, its pharmacological mechanism of action, common side effects and important interactions. In addition, they need to at least be aware of the evidence-base and trial data underlying its use, either first-hand or by following derived 'guidelines'. However, the pros and cons of taking drugs are much less appreciated by prescribers when it comes to individual patients. As a rule, both clinicians and patients tend to overestimate the benefits and underestimate the harms of medicines.[19] More 'real-life' research and studies are needed to inform prescribers.[20] In the future, it is likely that artificial intelligence generated algorithms, including pharmacogenetic variables, will also support this process.

Teaching and education around the practicalities of prescribing, adherence and polypharmacy need to be incorporated into medical school curricula as a matter of routine. The advent of the Prescribing Safety Assessment in the UK has certainly helped in this regard by focusing teaching on a hitherto neglected area.[21]

Furthermore, communication skills training for clinicians has tended to centre on explaining diagnoses, rather than the drugs patients take—why they need them, what they do, what effects the patient might feel, and what to note or look out for—all outlined in simple to understand language or graphics.

A single prescriber to co-ordinate all therapies, as opposed to having multiple prescribers, has obvious benefits in terms of making communication about drugs more efficient. Similarly, sharing the decision-making process with individual patients before starting or indeed stopping treatments is pivotal (figure 1).[22] Making time for this is demanding, particularly if the patient has cognitive

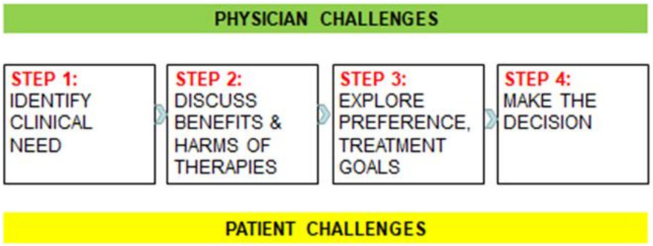

PHYSICIAN CHALLENGES
STEP 1: IDENTIFY CLINICAL NEED
STEP 2: DISCUSS BENEFITS & HARMS OF THERAPIES
STEP 3: EXPLORE PREFERENCE, TREATMENT GOALS
STEP 4: MAKE THE DECISION
PATIENT CHALLENGES

Figure 1 Decision-making challenges for physicians and patients.[22]

impairment or learning difficulties and the clinician has a finite consultation period. Agreeing on the 'goal of treatment' also presents a substantial hurdle: should the emphasis be primarily on longevity or on comfort and symptom control only? Finally, education on geriatric pharmacotherapy is needed to improve prescribing on geriatric patients.

Empowering patients by educating them about the medications they take is also a crucial element of ensuring maximal adherence.[23] Even basic steps such as encouraging patients themselves to keep an up-to-date list of their medications (prescribed, over the counter and herbal remedies) has evident advantages, especially during the transition between primary and secondary healthcare or vice versa.

As the coordination of patients' medication is often missing, explaining to patients how to monitor the effects of pharmacotherapy and identify potential risk is important.

The act of medicines reconciliation with pharmacist input is in itself an informative exercise. Many health IT tools such as integrated electronic prescribing platforms, and apps on mobile devices help to improve communication about drug prescribing and adherence to medication, respectively.[24] Most apps are based around 'reminder technology', although more sophisticated ones that help patients with their drugs list are becoming available. In this regard, the utilisation of dosette boxes and medication administration records for those with dementia and in care settings provides an excellent support tool.

Finally, regular review of medication lists is mandatory to safeguard against polypharmacy and maintaining the focus on the goals of therapy. Again, time and training to do this as well as building it into routine clinical practice is an increasing necessity.

IMPROVING ADHERENCE TO MEDICATION

Non-adherence to disease-modifying medications is an avoidable cause of emergency department (ED) visits and hospitalisations. The prevalence of non-adherence varies by condition and study[24–29] from 40% to 60% in chronic obstructive pulmonary disease,[24 25] 26% to 65% in myocardial infarction[26–28] and up to 93% in heart failure.[29] The estimated risk of ED visits and hospitalisations associated with non-adherence varies from 45% to 85%[30–32] and may be higher in patients with heart failure—a twofold increase in the risk of hospitalisation or death when adherence to disease-modifying medication is less than 80%.[33]

Previous studies reveal three primary reasons for medication non-adherence[34–38]:

- Cost.
- Fear of or experience with adverse medication effects.
- Ambivalence, or lack of perceived need or relevance to the patient.

Depression, other psychiatric problems and cognitive impairment also contribute to non-adherence[35 37 38] as do polypharmacy and complex drug regimens.[37 39]

Social support and a collaborative trusting relationship with the healthcare team increase adherence.

Systematic reviews of adherence interventions illustrate the wide variability in interventions evaluated and populations, conditions and medications targeted.[24 40–43] Almost all targeted single groups of drugs demonstrated modest effect sizes at best and only a minority improved clinical outcomes. Moreover, while most adherence interventions are multifaceted, they typically combine generic medication information with simplistic behavioural strategies (eg, reminders, pill organisers).[24 40–43] The Information-Motivation-Behavioural Skills model provides a comprehensive theoretical framework for designing adherence interventions, bringing together key aspects of behaviour change theories to target and improve adherence. Interventions that incorporate self-determination theory[38 44] and motivational interviewing[45–47] to directly target intrinsic motivation, confidence and autonomy have proven efficacy in smoking cessation,[48] weight loss[44] and medication adherence.[49] However, such interventions to change behaviour have not been extensively implemented due to resource intensity, inadequate health professional training and lack of reimbursement models to support implementation.[40 41]

Mobile technologies have emerged as popular and potentially powerful tools to provide individualised support to change health behaviours.[50] In 2019, 53% of older adults in the USA owned a smartphone.[51] This creates a new opportunity to assess how well mobile apps can improve health behaviours and outcomes among older adults. Although there are more than 700 medication management apps, most have not been evaluated,[52 53] and none have exploited the potential of this medium, limiting features to maintaining a medication list (by manual entry), pill reminders and refill requests.[52 54–57]

Hybrid interventions that combine mobile apps with monitoring and triage to the health team based on patient need can empower and motivate patients and caregivers via tools that help identify and address ambivalence, promote adaptive problem solving and provide quicker access to the health team to address knowledge gaps. Systematic reviews of web-based and hybrid interventions show that they increase patient empowerment, motivation, medication self-efficacy,[58–61] and in some cases patient outcomes.[60–64] To date, hybrid interventions have not been used to improve medication adherence.[59 61–63]

CONCLUSION

Better integrated digital and IT systems, improved education on prescribing for prescribers, and greater patient-centred care that empowers patients to take control of their medications are all vital to safer and more effective prescribing. Future research should leverage the considerable investment made by many countries in advancing digital healthcare infrastructures and develop and evaluate multifaceted hybrid interventions to reduce avoidable adverse events and improve adherence.

Funding The roundtable discussion was supported by an unrestricted educational grant by BD.

Competing interests DWB grants and personal fees from EarlySense, personal fees from CDI Negev, equity from ValeraHealth, equity from Clew, equity from MDClone, personal fees and equity from AESOP, personal fees and equity from Feelbetter, and grants from IBM Watson Health, outside the submitted work. H-YC None to declare. NTC None to declare. RKJ ECRI (ISMP is an affiliate of ECRI) has a non-binding agreement with AESOPFAM None to declare. RT None to declare.

Patient consent for publication Not applicable.

Provenance and peer review Commissioned; externally peer reviewed.

Data availability statement No data are available. Not applicable.

ORCID iD

Yu-Chuan Li http://orcid.org/0000-0001-6497-4232

REFERENCES

1 Dornan T, Ashcroft DM, Heathfield H. An in-depth investigation into causes of prescribing errors by Foundation trainees in relation to their medical education. EQUIP study. final report, 2009. Available: https://www.gmc-uk.org//media/documents/FINAL_Report_prevalence_and_causes_of_prescribing_errors.pdf_28935150.pdf

2 Krotki K, Henripin J. Baby boom, 2013. Available: https://www.thecanadianencyclopedia.ca/en/article/baby-boom [Accessed 8 Nov 2021].

3 Slawomirski L, Auraaen A, Klazinga N. *The economics of patient safety: strengthening a value-based approach to reducing patient harm at national level*. OECD Health Working Papers, 2017.

4 Call to Action. *Preventable health care harm is a public health crisis and patient safety requires a coordinated public health response*. Boston, MA: National Patient Safety Foundation, 2017.

5 Berner ES, Detmer DE, Simborg D. Will the wave finally break? a brief view of the adoption of electronic medical records in the United States. *J Am Med Inform Assoc* 2005;12:3–7.

6 Bates DW, Teich JM, Lee J, *et al*. The impact of computerized physician order entry on medication error prevention. *J Am Med Inform Assoc* 1999;6:313–21.

7 Kaushal R, Shojania KG, Bates DW. Effects of computerized physician order entry and clinical decision support systems on medication safety: a systematic review. *Arch Intern Med* 2003;163:1409.

8 Goundrey-Smith SJ. Technologies that transform: digital solutions for optimising medicines use in the NHS. *BMJ Health Care Inform* 2019;26:e100016.

9 Wright A, Aaron S, Seger DL, *et al*. Reduced effectiveness of Interruptive drug-drug interaction alerts after conversion to a commercial electronic health record. *J Gen Intern Med* 2018;33:1868–76.

10 Shah SN, Amato MG, Garlo KG, *et al*. Renal medication-related clinical decision support (CDS) alerts and overrides in the inpatient setting following implementation of a commercial electronic health record: implications for designing more effective alerts. *J Am Med Inform Assoc* 2021;28:1081–7.

11 Edrees H, Amato MG, Wong A, *et al*. High-priority drug-drug interaction clinical decision support overrides in a newly implemented commercial computerized provider order-entry system: override appropriateness and adverse drug events. *J Am Med Inform Assoc* 2020;27:893–900.

12 Ash JS, Sittig DF, Poon EG, *et al*. The extent and importance of unintended consequences related to computerized provider order entry. *J Am Med Inform Assoc* 2007;14:415–23.

13 Koppel R, Metlay JP, Cohen A, *et al*. Role of computerized physician order entry systems in facilitating medication errors. *JAMA* 2005;293:1197–203.

14 Melnick ER, Harry E, Sinsky CA, *et al*. Perceived electronic health record usability as a predictor of task load and burnout among US physicians: mediation analysis. *J Med Internet Res* 2020;22:e23382.

15 Salyers MP, Bonfils KA, Luther L, *et al*. The relationship between professional burnout and quality and safety in healthcare: a meta-analysis. *J Gen Intern Med* 2017;32:475–82.

16 Allianz global corporate and specialty. global aviation safety study: a review of 60 years of improvement in aviation safety, 2014. Available: https://www.agcs.allianz.com/content/dam/onemarketing/agcs/agcs/reports/AGCS-Global-Aviation-Safety-2014-report.pdf

17 Trinkley KE, Blakeslee WW, Matlock DD, *et al*. Clinician preferences for computerised clinical decision support for medications in primary care: a focus group study. *BMJ Health Care Inform* 2019;26:e000015.

18 Chernoby K, Lucey MF, Hartner CL, *et al*. Impact of a clinical decision support tool targeting QT-prolonging medications. *Am J Health Syst Pharm* 2020;77:S111–7.

19 Yang F, Wittes J, Pitt B. Beware of on-treatment safety analyses. *Clin Trials* 2019;16:63–70.

20 Krauss A. Why all randomised controlled trials produce biased results. *Ann Med* 2018;50:312–22.

21 PSA. Available: https://prescribingsafetyassessment.ac.uk/

22 Jansen J, Naganathan V, Carter SM, *et al*. Too much medicine in older people? deprescribing through shared decision making. *BMJ* 2016;353:i2893.

23 Comín-Colet J, Enjuanes C, Verdú-Rotellar JM, *et al*. Impact on clinical events and healthcare costs of adding telemedicine to multidisciplinary disease management programmes for heart failure: results of a randomized controlled trial. *J Telemed Telecare* 2016;22:282–95.

24 Nieuwlaat R, Wilczynski N, Navarro T, *et al*. Interventions for enhancing medication adherence. *Cochrane Database Syst Rev* 2014;11:Cd000011.

25 Restrepo RD, Alvarez MT, Wittnebel LD, *et al*. Medication adherence issues in patients treated for COPD. *Int J Chron Obstruct Pulmon Dis* 2008;3:371–84.

26 Bryant J, McDonald VM, Boyes A, *et al*. Improving medication adherence in chronic obstructive pulmonary disease: a systematic review. *Respir Res* 2013;14:109.

27 Tuppin P, Neumann A, Danchin N, *et al*. Evidence-based pharmacotherapy after myocardial infarction in France: adherence-associated factors and relationship with 30-month mortality and rehospitalization. *Arch Cardiovasc Dis* 2010;103:363–75.

28 Jackevicius CA, Li P, Tu JV. Prevalence, predictors, and outcomes of primary nonadherence after acute myocardial infarction. *Circulation* 2008;117:1028–36.

29 Choudhry NK, Setoguchi S, Levin R, *et al*. Trends in adherence to secondary prevention medications in elderly post-myocardial infarction patients. *Pharmacoepidemiol Drug Saf* 2008;17:1189–96.

30 Wu J-R, Moser DK, Lennie TA, *et al*. Medication adherence in patients who have heart failure: a review of the literature. *Nurs Clin North Am* 2008;43:133–53.

31 Shin S, Song H, Oh S-K, *et al*. Effect of antihypertensive medication adherence on hospitalization for cardiovascular disease and mortality in hypertensive patients. *Hypertens Res* 2013;36:1000–5.

32 Yang B, Pace P, Banahan B. Medication nonadherence and the risks of hospitalization, emergency department visits, and death among Medicare Part D enrollees with diabetes. *Patient Care* 2009;21.

33 Ho PM, Magid DJ, Shetterly SM, *et al*. Medication nonadherence is associated with a broad range of adverse outcomes in patients with coronary artery disease. *Am Heart J* 2008;155:772–9.

34 Fitzgerald AA, Powers JD, Ho PM, *et al*. Impact of medication nonadherence on hospitalizations and mortality in heart failure. *J Card Fail* 2011;17:664–9.

35 Gadkari AS, McHorney CA. Medication nonfulfillment rates and reasons: narrative systematic review. *Curr Med Res Opin* 2010;26:683–705.

36 Kennedy J, Tuleu I, Mackay K. Unfilled prescriptions of Medicare beneficiaries: prevalence, reasons, and types of medicines prescribed. *J Manag Care Pharm* 2008;14:553–60.

37 McHorney CA, Spain CV. Frequency of and reasons for medication non-fulfillment and non-persistence among American adults with chronic disease in 2008. *Health Expect* 2011;14:307–20.

38 Gellad WF, Grenard JL, Marcum ZA. A systematic review of barriers to medication adherence in the elderly: looking beyond cost and regimen complexity. *Am J Geriatr Pharmacother* 2011;9:11–23.

39 DiMatteo MR, Haskard-Zolnierek KB, Martin LR. Improving patient adherence: a three-factor model to guide practice. *Health Psychol Rev* 2012;6:74–91.

40 Vik SA, Maxwell CJ, Hogan DB. Measurement, correlates, and health outcomes of medication adherence among seniors. *Ann Pharmacother* 2004;38:303–12.

41 Demonceau J, Ruppar T, Kristanto P, *et al*. Identification and assessment of adherence-enhancing interventions in studies assessing medication adherence through electronically compiled drug dosing histories: a systematic literature review and meta-analysis. *Drugs* 2013;73:545–62.

42 Williams A, Manias E, Walker R. Interventions to improve medication adherence in people with multiple chronic conditions: a systematic review. *J Adv Nurs* 2008;63:132–43.

43 Conn VS, Ruppar TM, Enriquez M, *et al*. Medication adherence interventions that target subjects with adherence problems: systematic review and meta-analysis. *Res Social Adm Pharm* 2016;12:218–46.

44 Patton DE, Hughes CM, Cadogan CA, *et al*. Theory-Based interventions to improve medication adherence in older adults prescribed polypharmacy: a systematic review. *Drugs Aging* 2017;34:97–113.

45 Ryan R, Patrick H, Deci E. Facilitating health behaviour change and its maintenance: interventions based on self-determination theory. *Euro Health Psycholog* 2008;10.

46 Rollnick S, Miller WR. What is motivational interviewing? *Behav Cogn Psychother* 1995;23:325–34.

47 Hettema J, Steele J, Miller WR. Motivational interviewing. *Annu Rev Clin Psychol* 2005;1:91–111.

48 Rollnick S, Miller WR, Butler CC, *et al*. Motivational interviewing in health care: helping patients change behavior. *COPD* 2008;5:203.

49 Williams GC, McGregor HA, Sharp D, *et al*. Testing a self-determination theory intervention for motivating tobacco cessation: supporting autonomy and competence in a clinical trial. *Health Psychol* 2006;25:91–101.

50 Williams GC, Rodin GC, Ryan RM, *et al*. Autonomous regulation and long-term medication adherence in adult outpatients. *Health Psychol* 1998;17:269–76.

51 Free C, Phillips G, Galli L, *et al*. The effectiveness of mobile-health technology-based health behaviour change or disease management interventions for health care consumers: a systematic review. *PLoS Med* 2013;10:e1001362.

52 Anderson M. Mobile technology and home broadband 2019. *Pew Research Center* 2019.

53 Ahmed I, Ahmad NS, Ali S, *et al*. Medication adherence Apps: review and content analysis. *JMIR Mhealth Uhealth* 2018;6:e62.

54 Lancaster K, Abuzour A, Khaira M, *et al*. The use and effects of electronic health tools for patient self-monitoring and reporting of outcomes following medication use: systematic review. *J Med Internet Res* 2018;20:e294.

55 Haase J, Farris KB, Dorsch MP. Mobile applications to improve medication adherence. *Telemed J E Health* 2017;23:75–9.

56 Dayer L, Heldenbrand S, Anderson P, *et al*. Smartphone medication adherence apps: potential benefits to patients and providers. *J Am Pharm Assoc* 2013;53:172–81.

57 Heldenbrand S, Martin BC, Gubbins PO, *et al*. Assessment of medication adherence APP features, functionality, and health literacy level and the creation of a searchable web-based adherence APP resource for health care professionals and patients. *J Am Pharm Assoc* 2016;56:293–302.

58 Santo K, Richtering SS, Chalmers J, *et al*. Mobile phone Apps to improve medication adherence: a systematic stepwise process to identify high-quality Apps. *JMIR Mhealth Uhealth* 2016;4:e132.

59 Samoocha D, Bruinvels DJ, Elbers NA, *et al*. Effectiveness of web-based interventions on patient empowerment: a systematic review and meta-analysis. *J Med Internet Res* 2010;12:e23.

60 He T, Liu X, Li Y, *et al*. Remote home management for chronic kidney disease: a systematic review. *J Telemed Telecare* 2017;23:3–13.

61 Hale TM, Jethwani K, Kandola MS, *et al*. A remote medication monitoring system for chronic heart failure patients to reduce readmissions: a Two-Arm randomized pilot study. *J Med Internet Res* 2016;18:e91.

62 Hamine S, Gerth-Guyette E, Faulx D, *et al*. Impact of mHealth chronic disease management on treatment adherence and patient outcomes: a systematic review. *J Med Internet Res* 2015;17:e52.

63 Garabedian LF, Ross-Degnan D, Wharam JF. Mobile phone and smartphone technologies for diabetes care and self-management. *Curr Diab Rep* 2015;15:109.

64 Hanlon P, Daines L, Campbell C, *et al*. Telehealth interventions to support self-management of long-term conditions: a systematic Metareview of diabetes, heart failure, asthma, chronic obstructive pulmonary disease, and cancer. *J Med Internet Res* 2017;19:e172.

BMJ Health & Care Informatics

Development of a customised programme to standardise comorbidity diagnosis codes in a large-scale database

Robert C Osorio ,[1] Kunal P Raygor,[2] Adib A Abla[2]

To cite: Osorio RC, Raygor KP, Abla AA. Development of a customised programme to standardise comorbidity diagnosis codes in a large-scale database. *BMJ Health Care Inform* 2022;**29**:e100532. doi:10.1136/bmjhci-2021-100532

Received 15 December 2021
Accepted 09 April 2022

[1]School of Medicine, University of California San Francisco, San Francisco, California, USA
[2]Department of Neurological Surgery, University of California San Francisco, San Francisco, California, USA

Correspondence to
Robert C Osorio;
robert.osorio@ucsf.edu

ABSTRACT

Objectives The transition from ICD-9 to ICD-10 coding creates a data standardisation challenge for large-scale longitudinal research. We sought to develop a programme that automated this standardisation process.

Methods A programme was developed to standardise ICD-9 and ICD-10 terminology into one system. Code was improved to reduce runtime, and two iterations were tested on a joint ICD-9/ICD-10 database of 15.8 million patients.

Results Both programmes successfully standardised diagnostic terminology in the database. While the original programme updated 100 000 cells in 12.5 hours, the improved programme translated 3.1 million cells in 38 min.

Discussion While both programmes successfully translated ICD-related data into a standardised format, the original programme suffered from excessive runtimes. Code improvement with hash tables and parallelisation exponentially reduced these runtimes.

Conclusion Databases with ICD-9 and ICD-10 codes require terminology standardisation for analysis. By sharing our programme's implementation, we hope to assist other researchers in standardising their own databases.

INTRODUCTION

On 1 October 2015, the department of Health and Human Services updated the International Classification of Diseases (ICD) system by mandating the adoption of ICD-10 diagnosis codes in electronic medical records.[1] Serving as the new standard for naming and categorizing patient diagnoses, the ICD-10 system contains over five times more codes than ICD-9, posing a challenge for analysing longitudinal databases spanning both systems. Prior solutions have included the use of alternate coding systems, which are updated each time a new ICD system is released. Current literature is aimed at the accuracy and scope of these systems,[2 3] how they update with new ICD releases,[3 4] and how systems are similar or different.[5 6] These studies fail to address how to implement such a system on a large-scale database, where manual reference and cell-by-cell translation is infeasible. We sought to develop a programme that quickly and accurately standardises a dataset to one diagnostic coding system.

METHODS

A nationwide dataset of paediatric hospital discharges was examined. Originating from the Healthcare Cost and Utilisation Project (H-CUP), this Kids' Inpatient Database (KID) contained administrative data on 15.8 million hospital discharges across 2003–2016. The targets of our data manipulation were 20 columns of diagnosis codes that represented patient comorbidities at the time of surgery: while most cases in the database occurred during ICD-9's era, 3.1 million discharges (19.6%) occurred in the 2016 KID update, and thus had ICD-10 codes. As a solution to this difference, H-CUP offers Elixhauser Comorbidity Software, which assigns diagnosis names to comorbidities based on the ICD-9 or 10 system.[7] Prior to programme development and testing, we defined a successful programme as one which cross-referenced all ICD-10 codes to their corresponding comorbidity classification. The resulting database would contain all 15.8 million discharges using the same classification system.

Prior to development, a Microsoft Excel File was acquired from H-CUP, which listed ICD-10 diagnosis codes in the first column, and Elixhauser diagnosis names in the first row. The remaining cells were marked with a '1' if an ICD-10 code matched a corresponding comorbidity. This served as the 'dictionary' for our data translation. All computer code was developed and executed on RStudio, V.4.0.2.

A programme was written to examine each column of comorbidity data and extract any ICD-10 diagnosis codes encountered. Each code was individually compared with the 'dictionary': the programme scanned through rows until it found a matching ICD code, then scanned across that row until a '1' was seen (denoting it found a matching

Table 1 Computer program development, runtimes and relative efficiency

Program name	Time to complete 100 000 rows	Time to complete entire 3.1M translations	Relative efficiency
Linear programme	12.5 hours	16.1 days	1×
Parallelised programme with hash table	1.2 min	38 min	610.1×

A programme was developed that successfully standardised the comorbidity coding system used in a 15.7 million patient database spanning 2003–2016. Parallelising this programme and implementing a hash table increased the speed by more than 600-fold, allowing 3.1 million patient rows to be updated in under 40 min.

diagnosis). When a match was found, the column name (the diagnosis) was captured, and the corresponding column in the KID was marked as a '1,' denoting that patient as having this comorbidity. This process repeated until all diagnosis codes were translated in that patient row. The programme would then proceed to the next row in the database, and would start over on the new ICD-10 codes.

During development, code was tested on a random 1000 rows of data. Once it successfully translated these rows, the programme was deployed on the 3.1 million patients with ICD-10 codes. A duplicate of the programme was then created, and served as the starting point for runtime optimisation. In a similar fashion to the development of the original code, this new programme was tested on a random 1000 rows, then executed on the larger database.

RESULTS

Both programmes successfully translated ICD-10 codes to the Elixhauser comorbidity classification. Results on programme runtimes for the first iteration ('Linear') and the more efficient ('Parallelised') code are displayed in table 1. When testing runtimes for the linear code, it updated 100 000 rows in 12–13 hours, varying slightly in each test. As a result, this linear code would take 16 days to complete the 3.1 million target rows in our dataset. Programme testing was stopped after 7 days due to impracticality of runtime.

In development of a second iteration of code, runtime was reduced by targeting algorithm efficiency. Complexity was improved through conversion of the 'dictionary' into a hash table, exponentially reducing the number of computer operations performed. Runtime was further improved by breaking the data into subsets, and translating each subset simultaneously. On a computer with a 16-core processor, this allowed the 3.1 million discharges to be broken into 16 subsets of roughly 200 000 discharges. This parallelised code translated all 3.1 million rows in 38 min (1.2 min/100 000 samples), a more than 600-fold increase in processing speed compared with the original programme.

DISCUSSION

For longitudinal databases spanning across the 2010s, researchers face the challenge of analysing data that utilises both ICD-9 and ICD-10 codes. Prior literature addressed the creation and accuracy of standardised classification systems, but failed to discuss how to implement these systems on large databases where manual translation is impossible.[2–6] We successfully automated the standardisation of diagnostic terminology for a database of 15.8 million hospital discharges across 2003–2016. Databases of this size often pose a challenge for automated programmes, as evidenced by our initial programme's excessively long runtime. The subsequent programme we developed, however, ran more than 600 times faster, underscoring the significance of code quality in large scale data manipulation.

The largest gains in runtime can be attributed to the implementation of hash tables instead) of a 'dictionary' Excel file. When a computer iterates through an Excel dictionary of R rows and C columns, up to R * C comparisons are needed to find a match for just one comorbidity. When translating up to 20 comorbidities per row, for 3.1 million datapoints, these accumulate to roughly 62 million * R * C computer operations, guaranteeing excessive runtimes. A hash table is a data structure composed of a list of 'keys,' where each key is associated with one and only one 'value'. By converting our dictionary into a hash table with ICD-10 diagnosis codes as 'keys' and Elixhauser's comorbidity names as 'values,' translating diagnoses became exponentially simpler. Whereas the dictionary required R*C operations to find a match for a single ICD-10 code, a hash table requires just one action by the computer.

In addition to reducing programme complexity, code parallelisation also contributed to its faster runtimes. By splitting the data into 16 subsets to simultaneously translate, our programme ran 16 times faster. This parallelisation is possible due to multicore processors available in computers sold today.

Other advantages in the development of a customised programme include generalisability to future implementations. Our programme examines the number of processing cores on the computer running the algorithm, ensuring that data are always divided and analysed as efficiently as possible. Additionally, our programme should be easily implemented on any 'dictionary' that is plugged into our software, so that future systems such as ICD-11 may also be translated. Any 'dictionary' of reference values may be used, ensuring long-term utility of our algorithm in future practice of large-scale research.

CONCLUSION

Hash tables and parallelised code allowed us to standardise the coding system used by a 15.8 million patient database in under 40 min. We hope that by publishing our methods of translation on such a notably large database, we aid researchers in transforming other large datasets. When attempting to standardise data spanning multiple years, researchers should consider programming such as ours where hash tables and parallelisation allow extreme amounts of data review to be completed in an exponentially quicker time frame.

Contributors Conception and design: AA, RCO and KR. Acquisition of data: AA, RCO and KR. Analysis and interpretation of data: AA, RCO and KR. Drafting the article: RCO and KR. Critically revising the article: AA, RCO and KR. Reviewed submitted version of manuscript: AA, RCO and KR. Approved the final submission of the manuscript: AA. Overall study supervision: AA.

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.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data may be obtained from a third party and are not publicly available. Data are available through the Healthcare Cost and Utilisation Project.

ORCID iD

Robert C Osorio http://orcid.org/0000-0002-7669-2176

REFERENCES

1 The switch from ICD-9 to ICD-10: when and why. Available: https://icd.codes/articles/icd9-to-icd10-explained [Accessed 23 Oct 2021].

2 Feudtner C, Feinstein JA, Zhong W, *et al*. Pediatric complex chronic conditions classification system version 2: updated for ICD-10 and complex medical technology dependence and transplantation. *BMC Pediatr* 2014;14:199.

3 Glasheen WP, Cordier T, Gumpina R, *et al*. Charlson Comorbidity Index: *ICD-9* Update and *ICD-10* Translation. *Am Health Drug Benefits* 2019;12:188-197.

4 Glasheen WP, Renda A, Dong Y. Diabetes Complications Severity Index (DCSI)-Update and ICD-10 translation. *J Diabetes Complications* 2017;31:1007–13.

5 Hua-Gen Li M, Hutchinson A, Tacey M, *et al*. Reliability of comorbidity scores derived from administrative data in the tertiary hospital intensive care setting: a cross-sectional study. *BMJ Health Care Inform* 2019;26:e000016.

6 Brusselaers N, Lagergren J. The Charlson comorbidity index in registry-based research. *Methods Inf Med* 2017;56:401–6.

7 Elixhauser comorbidity software, version 3.7. Available: https://www.hcup-us.ahrq.gov/toolssoftware/comorbidity/comorbidity.jsp [Accessed 27 Oct 2021].

BMJ Health & Care Informatics

Operationalising fairness in medical AI adoption: detection of early Alzheimer's disease with 2D CNN

Luca Heising,[1,2] Spyros Angelopoulos [3]

To cite: Heising L, Angelopoulos S. Operationalising fairness in medical AI adoption: detection of early Alzheimer's disease with 2D CNN. *BMJ Health Care Inform* 2022;**29**:e100485. doi:10.1136/bmjhci-2021-100485

Received 05 October 2021
Accepted 09 January 2022

[1]Department of Radiation Oncology (Maastro), Maastricht University Medical Centre+, Maastricht, Netherlands
[2]Tilburg School of Economics and Management, Tilburg University, Tilburg, Netherlands
[3]Durham University Business School, Durham University, Durham, UK

Correspondence to
Dr Spyros Angelopoulos; Spyros.Angelopoulos@durham.ac.uk

ABSTRACT

Objectives To operationalise fairness in the adoption of medical artificial intelligence (AI) algorithms in terms of access to computational resources, the proposed approach is based on a two-dimensional (2D) convolutional neural networks (CNN), which provides a faster, cheaper and accurate-enough detection of early Alzheimer's disease (AD) and mild cognitive impairment (MCI), without the need for use of large training data sets or costly high-performance computing (HPC) infrastructures.
Methods The standardised Alzheimer's Disease Neuroimaging Initiative (ADNI) data sets are used for the proposed model, with additional skull stripping, using the Brain Extraction Tool V.2approach. The 2D CNN architecture is based on LeNet-5, the Leaky Rectified Linear Unit activation function and a Sigmoid function were used, and batch normalisation was added after every convolutional layer to stabilise the learning process. The model was optimised by manually tuning all its hyperparameters.
Results The model was evaluated in terms of accuracy, recall, precision and f1-score. The results demonstrate that the model predicted MCI with an accuracy of 0.735, passing the random guessing baseline of 0.521 and predicted AD with an accuracy of 0.837, passing the random guessing baseline of 0.536.
Discussion The proposed approach can assist clinicians in the early diagnosis of AD and MCI, with high-enough accuracy, based on relatively smaller data sets, and without the need of HPC infrastructures. Such an approach can alleviate disparities and operationalise fairness in the adoption of medical algorithms.
Conclusion Medical AI algorithms should not be focused solely on accuracy but should also be evaluated with respect to how they might impact disparities and operationalise fairness in their adoption.

Summary box

What is already known?

- Most prior studies on early Alzheimer's disease (AD) and mild cognitive impairment (MCI) detection have used a three-dimensional (3D) convolutional neural networks (CNN) approach.
- The 3D CNN approach is computationally expensive requiring high performance computing (HPC) infrastructures, and, due to the high number of parameters, it requires larger data sets for training.
- A two-dimensional (2D) CNN needs less parameters, less computational power and execution time, while requires smaller data sets for training, but has not been applied to date for MCI detection.

What does this paper add?

- The proposed approach based on a 2D CNN operationalises fairness in the adoption of medical artificial intelligence (AI) algorithms by providing fast, cheap and accurate-enough detection of early AD and MCI without the need for use of large data sets or costly HPC infrastructures.
- The proposed approach can be extended to other diseases as well as to other cases where time is scarce, powerful computational resources are not available, and large data sets are out of reach.

INTRODUCTION

Recent studies show that artificial intelligence (AI) applications can perform on par with medical experts on MRI analysis.[1] Such applications, to date, tend to oppose the accuracy of AI to the performance of clinicians. For instance, there have been more than 20 000 studies on deep learning (DL) methods for MRI analyses the last decade, which compare the performance of AI to the one of clinicians.[2] Recent work suggests that future studies should focus on the comparison of performance between clinicians using AI and their performance without an AI aid.[3] The recent global pandemic, however, revealed another urgent need of early disease diagnosis: the ability to make predictions based on a limited number of cases. The AI computer-aided detection (CAD) frameworks, to date, are based on large amounts of data and require high-performance computing (HPC) infrastructures. To address that lacuna, we propose a synergistic approach, in which clinicians and scientists collaborate for faster, cheaper and more accurate detection, relying on small data sets to make accurate-enough predictions. A promising frontier where AI can assist clinicians is Alzheimer's disease (AD) since the release of promising clinical studies for a new drug have unearthed the need for its early detection. As it can take up to 20 years before patients with AD show any signs of cognitive decline, it can be challenging to diagnose AD in early stages. We, thus, motivate and implement an AI-CAD framework for the early detection of mild cognitive impairment (MCI) and AD to assist clinicians, while the approach can be extended for the diagnosis of other diseases.

AD is caused by an accumulation of β-amyloid (Aβ) plaques, and abnormal amounts of *tau* proteins in the brain. This results in synapse loss, where the impulse does not reach the neurons, and in loss of structure or function of neurons, including their death, causing memory impairment and other cognitive problems.[4] AD has strong impact on the cognitive and physical functioning of patients, resulting in death. Recent developments in slowing AD decline have increased the relevance of its early detection,[5] and MCI plays an important role in this. MCI is a syndrome where the patients have greater cognitive decline than normally expected, but it does not necessarily affect their daily lives. Although some patients with MCI remain stable or return to cognitively normal (CN), there is a 10%–15% risk per year of progression to AD.[4] Before the aetiology of AD became known, its diagnosis relied on neurocognitive tests. The development of biomarkers improved AD detection. A common method to diagnose AD is hippocampus segmentation, which relates to memory function, and its small volume is an AD biomarker. For a long time, AD diagnosis was done manually by looking at the brain structure and size of the hippocampus on MRI, which requires practice and precision. Prior studies on automated methods for hippocampus segmentation have used DL approaches with promising results.[6] Automated hippocampus segmentation for the diagnosis of AD and MCI, however, requires clinicians' expertise and is sensitive to interrater and intrarater variability.[6]

Convolutional neural networks (CNN) can become the foundation of an AI-CAD framework for supporting clinicians in the detection of early AD and MCI, since it is a successful approach for image classification. CNN can improve the performance of image classification,[7] and they are becoming increasingly popular in MRI analysis. For instance, recent studies show that CNN can work on par with specialists for classifying MRI of patients with skin cancer.[1] Similar approaches with three-dimensional (3D) as well as two-dimensional (2D) CNN have also been used for AD detection with promising results. When it comes to the inner mechanics of these approaches, the classification filter of a 3D CNN slides along all the three dimensions of the input image, resulting in 3D feature maps, whereas in a 2D CNN the classification filter slides along only the height and width of the input image. Thus, the latter results in 2D feature maps, which need less parameters, computational power and execution time. Most prior studies have used 3D CNN achieving high accuracy,[8] while others obtained similar results with 2D CNN.[9] Although previous work on the topic has established that 3D CNN perform better for patch classifications, the results between 2D and 3D approaches for whole image labelling did not differ much.[10] A 3D CNN, however, is more computationally expensive, and, due to the high number of parameters, it requires larger data sets for training.[11] Concurrently, prior studies have not incorporated a 2D CNN approach for detecting MCI. A summary of prior 2D and 3D CNN applications in the literature is presented in table 1.

We suggest that medical algorithms should not be solely focused on accuracy but should also be evaluated with respect to how they might impact disparities and operationalise fairness in their adoption. Thus, we investigate the extent to which a 2D CNN can detect MCI and early AD.

METHODS

CNN is the most common neural network (NN) architecture for image classification. Fully connected NN take multiple inputs, and hidden layers perform calculations

Table 1 Performance comparison of 2D and 3D approaches in the literature

	2D CNN		3D CNN	
Study	**AD**	**MCI**	**AD**	**MCI**
Basaia *et al*[8]	–	–	0.99	0.87
Feng *et al*[16]	–	–	0.95	0.86
Korolev *et al*[25]	–	–	0.80	–
Liu *et al*[17]	–	–	0.85	–
Liu *et al*[18]	–	–	0.91	–
Senanayake *et al*[26]	–	–	0.76	0.75
Hon and Khan[27]	0.96	–	–	–
Sarraf and Tofighi[19]	0.99	–	–	–
Sarraf and Tofighi[20]	0.97	–	–	–
Wang *et al*[9]	0.98	–	–	–

AD, Alzheimer's disease; CNN, convolutional neural network; MCI, mild cognitive impairment.

on them, while the neurons in the network connect to each other. Neurons in CNN, however, connect only to those close to them. CNN, therefore, needs fewer parameters, which results in benefits such as small risk of overfitting, higher accuracy and faster processing time. Moreover, in CNN, there is no need to transform the input images to one dimensional, a process which can result in loss of structural information, as the CNN can learn the relationships among the pixels of input by extracting representative features with kernel convolutions[4] :

$$S\left(i,j\right) = \left(I \times K\right)\left(i,j\right) = \sum_{m}\sum_{n} I\left(m,n\right) K\left(i-m,j-n\right)$$

where I is the input and K is the kernel; the input indices are represented by i and j, and the kernel indices are represented by m and n.

The data sets used in this study were obtained under permission from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public–private partnership. The primary goal of ADNI has been to test whether MRI, biological markers and clinical as well as neuropsychological assessment can be combined to measure the progression of MCI and early AD. The ADNI is separated into three studies of 5 years, while the first was prolonged by 2 years under the name ADNI-GO. In total, 2517 people of ages 55–90 participated in the study. The ADNI encourages the use of their standardised data sets to ensure consistency in analysis and direct comparison of various methods among studies. We, therefore, used their two standardised data sets 'ADNI1: complete 2 year 1.5T' and 'ADNI1: complete 3 year 1.5T', which contain MRI that has passed quality control assessment.[12]

Our data set consists of 3312 images, distributed in 828 MRI of CN subjects, 453 MRI of patients with AD and 1203 MRI of patients with MCI. The data set was split into one with CN and AD subjects (1281 MRI), and one with CN and MCI subjects (2031 MRI). Since the participants of the ADNI study returned for more than one check-up, any patient can have up to 12 MRI, which are not identical as they are taken at different moments, and every MRI in the standardised data set was treated independently. The data set, thus, refers to 99 patients with AD, 212 patients with MCI and a control group of 165 CN subjects. We present the demographic information of the included subjects in table 2, to enable comparison with other studies.

Table 2 Demographic information of subjects in the dataset

	MCI	AD	CN
Images	891	412	662
Subjects	212	99	165
Gender	142 M / 70 F	52 M / 47 F	82 M / 83 F
Age	μ=75.84 σ=7.02	μ=76.49 σ=7.43	μ=76.93 σ=5.23

AD, Alzheimer's disease; CN, cognitively normal; MCI, mild cognitive impairment.

While the data sets are preprocessed, we further performed skull stripping using the Brain Extraction Tool V.2 (BET2), which is part of the *NiPype* library. Skull stripping locates the brain in the MRI and removes all surroundings to further remove noise from images. For optimal skull stripping results, neck slices were removed with the *robustfov* function. We used a fraction intensity of 0.3 as an evaluation of BET2 parameters for the ADNI data set found that this leads to best results. Due to the differences in scanners and techniques used by the ADNI over the years, the MRIs used in the data sets were of different sizes, and, therefore, had to become uniform. All the MRIs in our data set were resized to: (136, 192, 160) with the *ndimage* zoom function of the *Scipy* library, which zooms the array using spline interpolation. Resizing the MRI results in a different range of pixel values, and, therefore, to assure that the pixel values of all MRI had the same range, z-score normalisation was applied, which is defined as follows:

$$z_i = \frac{x_i - \mu\left(x\right)}{\sigma\left(x\right)}$$

where x is the MRI data and z_i the ith normalised MRI. The data set was then split into train set, validation set and test set with a ratio of 60:20:20, respectively.

An NN consists of an input layer, hidden layers and an output layer. A CNN has hidden layers divided into convolution, pooling, activation and classification layers. We based our architecture on LeNet-5, which includes two convolutional layers, two pooling layers and two fully connected layers (supplementary files, table 3).

We employ the Leaky Rectified Linear Unit (LReLU) as activation function for all convolutional layers, which allows for a small non-zero gradient.[13] The LReLU activation function in the model, with x being the input data, is described as:

$$\mathrm{y}\left(x\right) = \begin{cases} x, & if x < 0 \\ 0.01x, & otherwise \end{cases}$$

A Sigmoid activation function was applied to the dense layer, which outputs the probability of the images' class, with 0 if healthy and 1 if not (AD or MCI). The Sigmoid activation function in the model, with x being the input data, is described as:

$$\sigma\left(x\right) = \frac{1}{\left(1+e^{-x}\right)}.$$

We optimised the model by manually tunning the hyperparameters (see table 4).

The batch size was set to 16 and we used the Adam optimiser[14] with a learning rate of 10^{-3}. The model showed

Table 3 CNN architecture

Layer	C1	P1	C2	P2	FC1	FC2	FC3
Kernel	3×3	2×2	3×3	2×2	–	–	–
Filter	32	32	64	64	128	64	2

CNN, convolutional neural network.

Table 4 Parameter tuning on the AD dataset

Parameters											
Round	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10***	
Learning rate	0.0001	0.0001	0.001	0.001	0.001	0.001	0.001	0.0001	0.01	0.001	0.001
Batch size	32	16	16	8	32	16	8	8	8	8	16
Epochs	50	50	50	30	30	30	30	30	30	30	40
Dropout	–	0.3	0.3	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.2
Batch norm.	–	x	x	x	x	x	–	x	x	x	x
Metrics											
Loss	1.040	0.711	0.637	**0.461**	0.742	0.600	2.292	0.805	0.639	0.600	0.677
Acc	0.833	0.794	0.802	**0.840**	0.833	0.840	0.728	0.767	0.825	0.833	0.837
Precision	0.881	**0.977**	0.891	0.768	0.947	0.949	0.628	0.972	0.788	0.859	0.948
Recall	0.628	0.447	0.521	**0.809**	0.574	0.596	0.628	0.372	0.713	0.649	0.585

AD, Alzheimer's disease.

overfitting, which means that it includes more terms or uses more complicated approaches than necessary.[15] Regularisation can control overfitting and drop-out regularisation is a commonly used approach because it is computationally inexpensive, and it prevents coadaptation among feature map units.[11] In drop-out regularisation, only a fraction of the weights is learnt by the NN in each iteration. We added a drop-out layer with a value of 0.2 after each pooling layer (ie, 80% of the weights were learnt in each iteration), leading to better results on all the train, validation and test sets. To stabilise the learning process, we added batch normalisation after every convolutional layer. For each unit in a layer, the value was normalised as follows:

$$a_i^{(l)} = \frac{a_i^{(l)} - E\left[a_i^{(l)}\right]}{\sqrt{Var\left[a_i^{(l)}\right]}}$$

where a represents the activation vector of the *ith* layer l. Thereafter, the normalised values were scaled and shifted accordingly. After ~40 epochs, the model did not show increment in accuracy or reduction in loss, and overfitting increased, thus, we applied an early stopping at 40 epochs instead of the initial set of 50.

The CNN was built with a *Jupyter Notebook* using *Python V.3.6.4*, *Tensorflow V.2.4.0* and *Keras V.2.4.0*. To load the data in *NIfTI* format, we used the *Nilearn* library, and we used the *scikit-learn* and *SciPy* libraries for data preprocessing. The development, testing and application of the model took place on the Google Cloud Console, where we used a storage bucket to store the data sets, and three compute engine instances to perform the skull stripping and preprocessing and to run our model independently as these steps require different computational resources. For skull stripping, we used an instance with 8 vCPUs, 52 GB RAM, and two NVIDA Tesla K80 GPUs, for preprocessing, we used an instance with 40 vCPUs and 961 GB RAM. For the CNN, we used an instance with 64 vCPUs, 416 GB RAM and four NVIDA Tesla T4 GPUs.

RESULTS

The model was evaluated in terms of *accuracy*, *recall*, *precision* and *f1-score*. Recall provides sensitivity information on how many patients were correctly identified. Precision expresses how many of the positives that the model returns were actually positive. F1-score is the harmonic mean between precision and recall. An NN adjusts its weights to optimise the loss, which is calculated with the use of binary cross entropy loss:

$$CE = -\sum_{i=1}^{C'=2} t_i log\left(s_i\right) = t_1 log\left(s_1\right) - \left(1 - t_1\right) log\left(1 - s\right)$$

where C represents the classes, s_i is the predicted probability value for class i and t is the true probability for that class. Since the data were unevenly distributed, the accuracy baseline of random guessing was also calculated. The baseline was calculated with respect to the class distribution of the data set. First, we trained and tested our

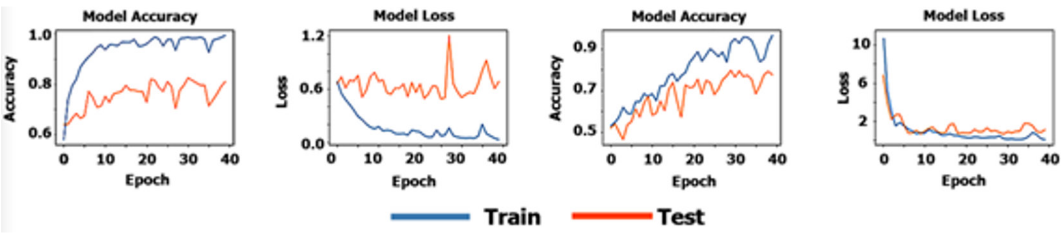


Figure 1 Model performance for the AD and MCI datasets. AD, Alzheimer's disease; MCI, mild cognitive impairment.

Table 5 Performance metrics on test data

Data	Loss	Accuracy	Precision	Recall	F1	MRI
AD	0.677	0.837	0.948	0.585	0.724	1281
MCI	1.377	0.735	0.728	0.894	0.802	2031

AD, Alzheimer's disease; MCI, mild cognitive impairment.

model on the AD data set. After passing the baseline of random guessing on the training data (>0.548) with an accuracy of 0.994, we applied the same model on the MCI data set. The random guessing baseline for the test data set of the AD model was 0.536 and for the test data set of the MCI model was 0.521. The overepochs performance of the model is depicted in figure 1 for AD (left) and for MCI (right).

While the above graphs indicate a normal learning curve, as the performance of the model keeps increasing on the train data set, the validation performance flattens, which implies overfitting. This appears to be true mainly on the AD data set. Our model achieved accuracy of 0.837 on the AD test set. Irrespective of overfitting, the achieved test accuracy on the AD data set surpasses the random guessing baseline of 0.536. The model predicted MCI with accuracy of 0.735, passing the random guessing baseline of 0.521. Table 5 presents the performance metrics of the models on the test sets. The model performs better than chance on both sets, with a better predictive performance for the AD data set than for the MCI data set. The MCI model, however, seems to perform better on selecting relevant items (ie, recall, predicted positives relative to all positives). The MCI model shows notably less overfitting than the AD model, which might be due to the size of the data set, as the dataset used for the MCI was larger (almost double in size) than the AD one.

By comparing our study to previous ones in the relevant literature (see table 6), we notice a large difference in the size of the used data sets. Moreover, some of the prior studies only report the number of subjects in the used data set,[8 16–18] but the number of images can differ from these since one subject can have up to 12 images in these data sets. As expected, studies with larger data sets achieved higher accuracy. Furthermore, some of the studies with a 2D approach treated the slices independently,[9 19 20] thereby enlarging the size of their data set, however, the MRI was not treated as a whole.

DISCUSSION

While AI-CAD frameworks have been thoroughly studied, they have not been proposed as a tool for assisting clinicians. Furthermore, while the literature on AI-CAD frameworks is mostly approached from a computer science perspective, clinicians have been shown to lack trust in them.[2 3 21] Our work addresses that lacuna by providing a synergistic approach between clinicians and scientists. We contribute to the line of research on using CNN for AD and MCI detection, by applying a 2D approach. Our model predicts AD better than chance by 0.301 and MCI by 0.214. As expected, the model performed worse on detecting MCI than AD. The learning process on the MCI data set, however, was much cleaner than the process on the AD data set. This might be due to the size of the data set, which can have a large impact on the process and outcomes of the model. The proposed AI-CAD framework, thus, performs better than chance for AD as well as for MCI and could assist clinicians in the early detection of AD and MCI.

We suggest that medical algorithms should not be focused solely on accuracy but should also be evaluated with respect to how they might impact disparities and operationalise fairness in terms of computational resources, when it comes to their adoption. Our framework can be

Table 6 Comparison of data and accuracy with previous studies

				Accuracy	
Study	Subjects	Images	Dimensions	AD	MCI
Basaia *et al*[8]	645	–	3D	0.99	0.87
Feng *et al*[16]	193	–	3D	0.95	0.86
Korolev *et al*[25]	111	111	3D	0.80	–
Liu *et al*[17]	193	–	3D	0.85	–
Liu *et al*[18]	902	–	3D	0.91	–
Senanayake *et al*[26]	–	322	3D	0.76	0.75
Hon and Khan*[27]	200	6400	2D	0.96	–
Sarraf and Tofighi**[19]	302	62 335	2D	0.99	–
Sarraf and Tofighi**[20]	43	367 200	2D	0.97	–
Wang *et al*.**[9]	98	17 738	2D	0.98	–
Our	476	3312	2D	0.84	0.74

*Accuracy before transfer learning=0.74.
†Used MRI slices independently.
AD, Alzheimer's disease; MCI, mild cognitive impairment.

further extended to other diseases, and to cases where time is scarce, computational resources are not available, and large data sets are out of reach. Finally, our work is in line with the broader Information Systems research agenda,[22] on the adoption of responsible medical AI algorithms,[23] and the stewardship of sensitive personal data.[24] Therefore, our work can give rise to new avenues for interdisciplinary research and can become the bedrock for novel methodological advances as well as groundbreaking empirical findings on the broader topic.

CONCLUSION

Prior studies have used CNN to diagnose MCI and early AD, most of which applied 3D approached. The 3D CNN, however, have drawbacks that relate to needs for HPC infrastructures. Other studies have focused on detecting AD with a 2D CNN, achieving similar results as the 3D approach. Despite the relevance of detecting MCI, prior studies did not investigate how these methods perform on detecting MCI. Our main goal was to determine whether a 2D CNN can be used to diagnose AD and MCI. Our work resulted in an AI-CAD framework that can assist clinicians in the early detection of MCI and AD with high-enough accuracy, based on a relatively small data set, and without the need of HPC infrastructures. Our work has limitations that need to be acknowledged. First, an important preprogressing step is image resizing. We used *Scipy ndimage*, which distorts the image and could have a negative effect on the learning process. A better solution for resizing images is needed but to the best of our knowledge is not available. Second, the ADNI data sets consist of more images than participants. If subjects appear in both data sets, the model could learn subject-specific features, but the impact on model performance is unknown, as most physical features are removed during skull stripping. Third, the AD model appears to be overfitting, which is a common problem in DL models. To further optimise our model, the overfitting problem needs to be addressed by future research. Future research should also replicate the existing 3D CNN approaches and compare their execution time with the 2D CNN one of our models on the same computational infrastructure. Such a comparison will further illustrate the merits of our approach. Finally, future research should also evaluate the performance of clinicians using our framework and their performance without an AI aid.

Twitter Spyros Angelopoulos @angelopoulos

Contributors LMH conceived the idea for the project and SA oversaw its overall direction and planning. Both LMH and SA worked on the acquisition of the data for analysis. LMH designed the computational framework and analysed the data. SA worked out the technical details for analysis and helped in the interpretation of the results. LMH is the guarantor of the study.

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 Durham University Business School, DUBS-2021-10-31T21:05:04-xggj42.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data was obtained from an open access database, which is publicly available. Not applicable.

ORCID iD

Spyros Angelopoulos http://orcid.org/0000-0002-8165-8204

REFERENCES

1 Esteva A, Kuprel B, Novoa RA, *et al*. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017;542:115–8.
2 Liu X, Faes L, Kale AU, *et al*. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Digit Health* 2019;1:e271–97.
3 Briganti G, Le Moine O. Artificial intelligence in medicine: today and tomorrow. *Front Med* 2020;7:27.
4 Weiner MW, Veitch DP, Aisen PS, *et al*. 2014 update of the Alzheimer's disease neuroimaging Initiative: a review of papers published since its inception. *Alzheimers Dement* 2015;11:e1–120.
5 Selkoe DJ. Alzheimer disease and aducanumab: adjusting our approach. *Nat Rev Neurol* 2019;15:365–6.
6 Ataloglou D, Dimou A, Zarpalas D, *et al*. Fast and precise hippocampus segmentation through deep Convolutional neural network ensembles and transfer learning. *Neuroinformatics* 2019;17:563–82.
7 Li Q, Cai W, Wang X. Medical image classification with convolutional neural network. *Proc Int Conf Control Automation Robotics Vision* 2014:844–8.
8 Basaia S, Agosta F, Wagner L, *et al*. Automated classification of Alzheimer's disease and mild cognitive impairment using a single MRI and deep neural networks. *Neuroimage Clin* 2019;21:101645.
9 Wang S-H, Phillips P, Sui Y, *et al*. Classification of Alzheimer's disease based on eight-layer convolutional neural network with leaky rectified linear unit and max pooling. *J Med Syst* 2018;42:85.
10 Lai M. Deep learning for medical image segmentation. *ArXiv* 2015.
11 Bernal J, Kushibar K, Asfaw DS, *et al*. Deep convolutional neural networks for brain image analysis on magnetic resonance imaging: a review. *Artif Intell Med* 2019;95:64–81.
12 Wyman BT, Harvey DJ, Crawford K, *et al*. Standardization of analysis sets for reporting results from ADNI MRI data. *Alzheimers Dement* 2013;9:332–7.
13 Lu L, Shin Y, Su Y. Dying ReLU and initialization: theory and numerical examples. *ArXiv* 2019.
14 Kingma DP, Ba J. Adam: a method for stochastic optimization. *ArXiv* 2014.
15 Hawkins DM. The problem of overfitting. *J Chem Inf Comput Sci* 2004;44:1–12.
16 Feng C, Elazab A, Yang P, *et al*. Deep Learning Framework for Alzheimer's Disease Diagnosis via 3D-CNN and FSBi-LSTM. *IEEE Access* 2019;7:63605–18.
17 Liu M, Cheng D, Wang K, *et al*. Multi-Modality Cascaded Convolutional neural networks for Alzheimer's disease diagnosis. *Neuroinformatics* 2018;16:295–308.
18 Liu M, Zhang J, Adeli E, *et al*. Landmark-based deep multi-instance learning for brain disease diagnosis. *Med Image Anal* 2018;43:157–68.
19 Sarraf A, Tofighi G. DeepAD: Alzheimer's Disease Classification via Deep Convolutional Neural Networks using MRI and fMRI. *BioRxiv* 2016:070441.
20 Sarraf A, Tofighi G. Classification of alzheimer's disease using fMRI data and deep learning convolutional neural networks. *ArXiv* 2016.
21 Markus AF, Kors JA, Rijnbeek PR. The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. *J Biomed Inform* 2021;113:103655.
22 Struijk M, CXJ O, Davison RM, *et al*. Putting the is back into is research. *Inf Syst J* 2022;32.

23 Trocin C, Mikalef P, Papamitsiou Z. Responsible AI for digital health: a synthesis and a research agenda. *Inf Syst Front* 2021:1–19.

24 Angelopoulos S, Brown M, McAuley D, *et al*. Stewardship of personal data on social networking sites. *Int J Inf Manage* 2021;56:102208.

25 Korolev S, Safiullin A, Belyaev M. Residual and plain convolutional neural networks for 3D brain MRI classification. *Proc IEEE Int Symp Biomed Imaging* 2017:835–8.

26 Senanayake U, Sowmya A, Dawes L. Deep fusion pipeline for mild cognitive impairment diagnosis. *Proc IEEE Int Symp Biomed Imaging* 2018:1394–997.

27 Hon M, Khan NM. Towards Alzheimer's disease classification through transfer learning. *Proc IEEE Int Conf Bioinformatics Biomed* 2017:1166–9.

BMJ Health & Care Informatics

Evaluating algorithmic fairness in the presence of clinical guidelines: the case of atherosclerotic cardiovascular disease risk estimation

Agata Foryciarz ,[1,2] Stephen R Pfohl,[2] Birju Patel,[2] Nigam Shah [2]

To cite: Foryciarz A, Pfohl SR, Patel B, *et al*. Evaluating algorithmic fairness in the presence of clinical guidelines: the case of atherosclerotic cardiovascular disease risk estimation. *BMJ Health Care Inform* 2022;**29**:e100460. doi:10.1136/bmjhci-2021-100460

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bmjhci-2021-100460).

Received 31 July 2021
Accepted 17 December 2021

[1]Department of Computer Science, Stanford University School of Engineering, Stanford, California, USA
[2]Stanford Center for Biomedical Informatics Research, Stanford University School of Medicine, Stanford, California, USA

Correspondence to
Agata Foryciarz;
agataf@stanford.edu

ABSTRACT

Objectives The American College of Cardiology and the American Heart Association guidelines on primary prevention of atherosclerotic cardiovascular disease (ASCVD) recommend using 10-year ASCVD risk estimation models to initiate statin treatment. For guideline-concordant decision-making, risk estimates need to be calibrated. However, existing models are often miscalibrated for race, ethnicity and sex based subgroups. This study evaluates two algorithmic fairness approaches to adjust the risk estimators (group recalibration and equalised odds) for their compatibility with the assumptions underpinning the guidelines' decision rules. MethodsUsing an updated pooled cohorts data set, we derive unconstrained, group-recalibrated and equalised odds-constrained versions of the 10-year ASCVD risk estimators, and compare their calibration at guideline-concordant decision thresholds.

Results We find that, compared with the unconstrained model, group-recalibration improves calibration at one of the relevant thresholds for each group, but exacerbates differences in false positive and false negative rates between groups. An equalised odds constraint, meant to equalise error rates across groups, does so by miscalibrating the model overall and at relevant decision thresholds.

Discussion Hence, because of induced miscalibration, decisions guided by risk estimators learned with an equalised odds fairness constraint are not concordant with existing guidelines. Conversely, recalibrating the model separately for each group can increase guideline compatibility, while increasing intergroup differences in error rates. As such, comparisons of error rates across groups can be misleading when guidelines recommend treating at fixed decision thresholds.

Conclusion The illustrated tradeoffs between satisfying a fairness criterion and retaining guideline compatibility underscore the need to evaluate models in the context of downstream interventions.

Summary

What is already known?

- Algorithmic fairness methods can be used to quantify and correct for differences in specific model performance metrics across groups, but the choice of an appropriate fairness metric is difficult.
- The pooled cohort equations (PCEs), 10-year atherosclerotic cardiovascular disease risk prediction models used to guide statin treatment decisions in the USA, exhibit differences in calibration and discrimination across demographic groups, which can lead to inappropriate or misinformed treatment decisions for some groups.
- Two theoretically incompatible fairness adjustments have been separately proposed for re-deriving the PCEs.

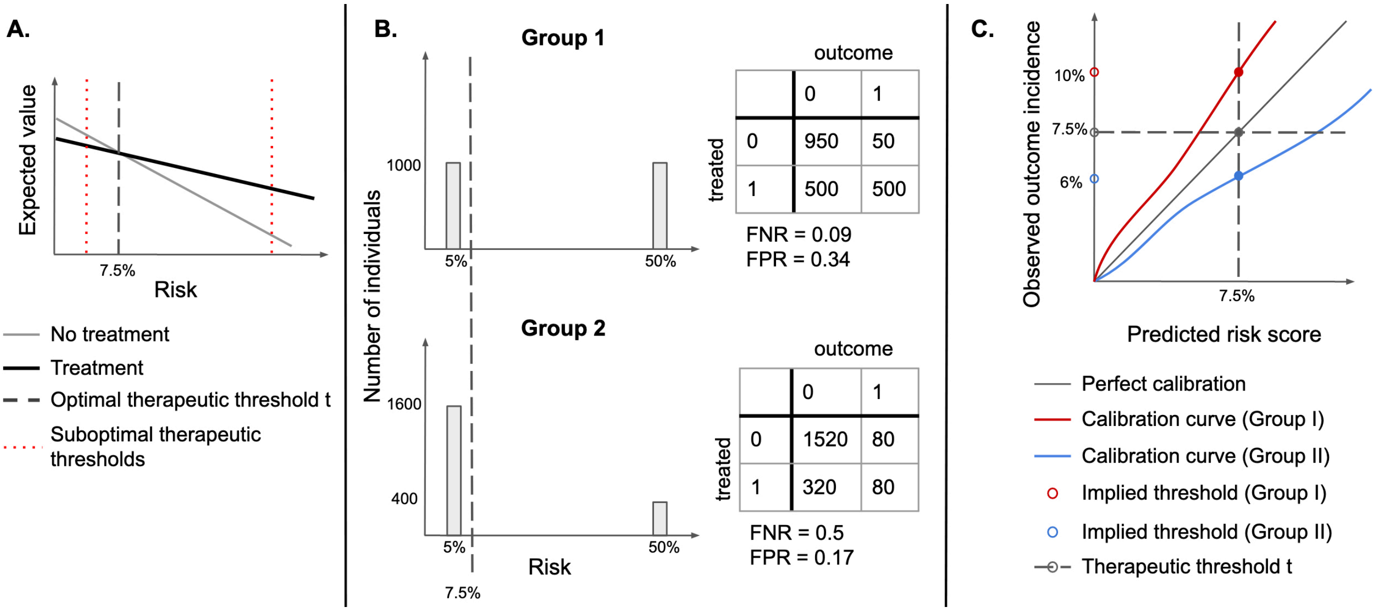
What does this paper add?

- Proposes a measure of local calibration of the PCEs at therapeutic thresholds as a method for probing guideline compatibility.
- Quantifies the effect of two proposed fairness methods for re-deriving the PCEs in terms of their impact on local calibration.
- Illustrates general principles that can be used to conduct contextually-relevant fairness evaluations of models used in clinical settings in the presence of clinical guidelines.

INTRODUCTION

While risk stratification models are central to personalising care, their use can worsen health inequities.[1] In an effort to mitigate harms, several recent works propose *algorithmic group fairness*—mathematical criteria which require that certain statistical properties of a model's predictions not differ across groups.[2 3] However, identifying which statistical properties are most relevant to fairness in a given context is non-trivial. Hence, before applying fairness criteria for evaluation or model adjustment, it is crucial to examine how the model's predictions will inform treatment decisions—and what effect those decisions will have on patients' health.

Here, we consider the 2019 guidelines of the American College of Cardiology and the American Heart Association (ACC/AHA) on primary prevention of atherosclerotic cardiovascular disease (ASCVD),[4] which codify



Group 1

treated \ outcome	0	1
0	950	50
1	500	500

FNR = 0.09
FPR = 0.34

Group 2

treated \ outcome	0	1
0	1520	80
1	320	80

FNR = 0.5
FPR = 0.17

Figure 1 (A) Identifying an optimal therapeutic threshold. An individual with risk r should be treated if the expected value of treatment exceeds that of non-treatment. As risk increases, the benefits of treatment become more significant, and assigning treatment becomes more optimal than withholding it. The optimal therapeutic threshold t is the value of risk at which treatment and non-treatment have the same expected value (the indifference point)—for individuals with r>t, treatment is expected to be more beneficial than non-treatment. Setting a non-optimal therapeutic threshold could lead to suboptimal treatment decisions for some individuals (treating some individuals for whom non-treatment has a higher expected value, or not treating individuals for whom treatment has a higher expected value). (B) Illustration of the sensitivity of FPR and FNR to the distribution of risk. Assume that there are two types of easily distinguishable individuals: with 5% and 50% chance of developing a disease, respectively, and there are two groups composed of both types of individuals, but one has a higher proportion of lower-risk individuals. If the same therapeutic threshold is applied to both groups, false positive rates (FPR) and false negative rates (FNR) will not be equal, even though we would be making optimal treatment decisions for each patient, in both populations. (C) Under miscalibration, implied thresholds differ from therapeutic thresholds. If risk scores are miscalibrated, taking action at the threshold of 7.5% corresponds to different observed outcome rates in the two groups. For Group I, a risk score of 7.5% corresponds to an observed outcome incidence of 10%, while for Group II it corresponds to 6%, therefore, individuals in Group II would be treated at a lower risk than individuals in Group I.

the use of 10-year ASCVD risk predictions to inform a clinician-patient shared decision-making on initiating statin therapy. These guidelines recommend that individuals estimated to be at intermediate risk (>7.5%–20%) be considered for initiation for moderate-intensity to high-intensity statin therapy, and that those at high risk (>20%) be considered for high-intensity statin therapy. Individuals at borderline risk (>5%–7.5%) may be considered for therapy under some circumstances.[4 5]

These *therapeutic thresholds* were established based on randomised control trials, and correspond to risk levels where expected overall benefits derived from low-density lipoprotein cholesterol reduction outweigh risks of side effects (online supplemental file C).[4 6] In general, such thresholds can be identified using decision analysis methods[7] (figure 1A). The models accompanying the guidelines (pooled cohort equations, PCEs[6 8 9]), developed for Black women, white women, Black men and white men, differ in both calibration and discrimination across groups.[10 11] The resultant systematic bias in risk misestimation in these subgroups can lead to inappropriate or misinformed treatment decisions. Since then, several works derived updated equations,[11–14] some explicitly incorporating fairness adjustments.[13 14]

If therapeutic thresholds recommended by guidelines reflect a balance of relevant harms and benefits for all subgroups,[15 16] therapeutic decisions could be unfair if thresholds used for different groups differ, as they would lead to suboptimal treatment decisions for some groups (figure 1A). As such, subgroup calibration at optimal therapeutic thresholds is an important fairness criterion for 10-year ASCVD risk estimation models,[14] since under miscalibration (systematic overestimation or underestimation of risk), treatment thresholds implicitly change (figure 1C) from treatment thresholds to *implied thresholds*.[17 18]

An alternative fairness criterion, known as *equalised odds (EO)*,[3] which has previously been used to evaluate several clinical predictive models,[13 19 20] requires equality in false positive and false negative error rates (FPR and FNR) across groups. One work proposed to explicitly incorporate EO constraints into the training objective to learn ASCVD risk estimators with minimal intergroup differences in FPR and FNR.[13]

In the context of ASCVD risk estimation, the EO criterion lacks a clear motivation and can thus yield misleading results. FPR and FNR are sensitive to the distribution of risk and are expected to differ across groups when the incidence of outcomes differs (figure 1B).[18 21 22]

Furthermore, approaches to build EO-satisfying models either explicitly adjust group-specific decision thresholds, introduce differential miscalibration or reduce model fit for each group[3]—which may lead to suboptimal decisions (figure 1A,C). EO-satisfying models may therefore be less appropriate than calibrated estimators for use with the ACC/AHA guidelines.[17]

We aim to evaluate the tension between calibration, EO and guideline-concordant decision-making. To do so, we propose a measure of local calibration at guideline-concordant therapeutic thresholds as a method for probing guideline compatibility and apply it to unconstrained, group-recalibrated and EO-constrained versions of the 10-year ASCVD risk prediction models learnt from the updated pooled cohorts data set,[9 11] as well as the original[8] and revised PCEs.[11] We assess the proposed local calibration measure and error rates across groups for each model, and conclude with recommendations for identifying quantification and adjustment criteria for enabling fair model-guided decisions.

METHODS

Data sets

We use an updated pooled cohorts data set,[11] comprised of ARIC (Atherosclerosis Risk in Communities Study, 1987–2011), CARDIA (Coronary Artery Risk Development in Young Adults Study, 1983–2006), CHS (Cardiovascular Health Study, 1989–1999), FHS OS (Framingham Heart Study Offspring Cohort, 1971–2014), MESA (Multi-Ethnic Study of Atherosclerosis, 2000–2012) and JHS (Jackson Heart Study, 2000–2012). Following the original PCE inclusion criteria,[9] we include individuals aged 40–79, excluding those with a history of myocardial infarction, stroke, coronary bypass surgery, angioplasty, congestive heart failure or atrial fibrillation, or receiving statins at the time of the initial examination. We include all individuals, regardless of racial category, and classify them as Black and non-Black, consistent with the use of PCEs in practice for non-Black patients of colour.

We extract features included in the PCEs (total cholesterol, high-density lipoprotein (HDL) cholesterol, treated and untreated systolic blood pressure, diabetes, smoking status, age, binary sex and race) and body mass index, recorded at the initial examination. We also extract dates of observed ASCVD events (myocardial infarction, lethal or non-lethal stroke or lethal coronary heart disease), and of last recorded observation (follow-up or death), to define binary labels for 10-year ASCVD outcome and censoring. Individuals whose last recorded observation happened before an ASCVD event and before year 10 are considered censored. We remove records with extreme values of systolic blood pressure (outside 90–200 mm Hg), total cholesterol and HDL cholesterol (outside 130–320 and 20–100 mg/dL, respectively) or missing covariates.

Models

Unconstrained model

The original PCEs consist of four separate Cox proportional hazards models, stratified by sex and race, to account for differences in ASCVD incidence across the four groups (Black women, white women, Black men and white men).[8] One revision of the PCEs, which reduced overfitting and improved calibration, replaced the Cox models with censoring-adjusted logistic regression models, stratified by sex and included race as a variable in each model.[11] Our implementation of the unconstrained (UC) models consists of a single inverse probability of censoring (IPCW)-adjusted logistic regression model,[23] and includes race and sex as binary variables. Censoring weights are obtained from four group-level Kaplan-Meier estimators applied to the training set. We include all features and their two-way interactions.

Group-recalibrated model

For recalibration, we logit-transform the predicted probabilities generated by the UC model, and use IPCW-adjusted logistic regression to fit a calibration curve for each group. We then use the resulting group-recalibrated model to obtain a set of recalibrated predictions.

EO model

The EO criterion requires that both the FPR and FNR be equal across groups at one or more thresholds.[3] We use an in-processing method for constructing EO models,[22 24] which provides a better calibration-EO tradeoff than the post-processing approach.[25] We define the training objective by adding a regulariser to the UC model's objective (online supplemental file A), with the degree of regularisation controlled by λ. The regulariser penalises differences between FPR and FNR at specified decision thresholds (7.5% and 20%), across the four groups.

Training procedures

Using random sampling stratified by group, outcome and presence of censoring, we divide our cohort into the training (80%), recalibration (10%) and test (10%) sets. Using the same procedure, we divide the training set into 10 equally-sized subsets and, for each subset, train a logistic regression model using stochastic gradient descent for up to 200 iterations of 128 minibatches, with learning rate of 10^{-4} on the remaining subsets. We terminate training if the cross-entropy loss does not improve on the held-out subset for 30 iterations. This procedure generates 10 UC models. To generate group-recalibrated models, we first generate predictions on the recalibration set, using the UC models (figure 2) and then use those train logistic regression models using BFGS (Broyden-Fletcher-Goldfarb-Shanno) optimisation, implemented in Scikit-Learn,[26] with up to 10^5 iterations. To examine the impact of the EO penalty, we repeat the unconstrained training procedure using the regularised training objective with four different settings of the parameter λ, distributed log-uniformly on the interval 0.1–1.0 (0.100,

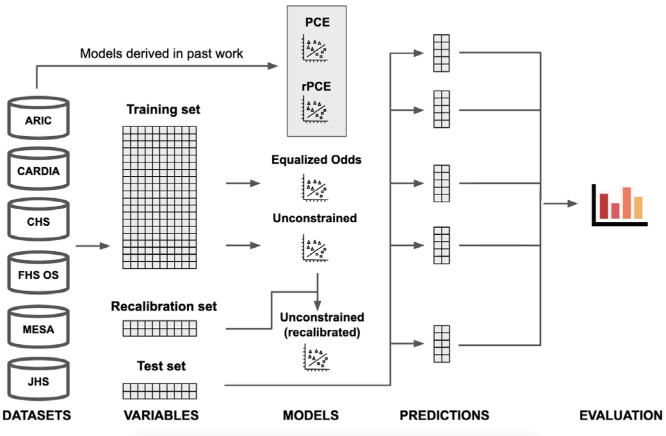


Figure 2 Visual abstract. Data from the six considered data sets: ARIC (Atherosclerosis Risk in Communities Study), CARDIA (Coronary Artery Risk Development in Young Adults Study), CHS (Cardiovascular Health Study), FHS OS (Framingham Heart Study Offspring Cohort), MESA (Multi-Ethnic Study of Atherosclerosis) and JHS (Jackson Heart Study), is extracted using the cohort definition used in the original pooled cohort equations (PCEs), and divided into train (80%), validation (10%) and test (10%) sets. Equalised odds and unconstrained (UC) models are derived directly from the training set. The recalibrated model is derived from the UC model using a recalibration procedure, which uses the validation data set (not seen during training). Finally, predictions on the test set are generated for all models—including the PCEs and the revised PCEs (rPCE), derived in past work—and evaluated.

0.215, 0.464, 1.000) and refer to the resulting models as EO1 through EO4. PyTorch V.1.5.0[27] is used to define all models and training procedures. We make our code available at https://github.com/agataf/fairness_eval_ascvd.

Evaluation

We introduce *threshold calibration error* (TCE), a measure of local calibration, defined as the difference between the therapeutic threshold (t_1=7.5% or t_2=20%) applied on the risk estimate and the *implied threshold* on the risk, measured by the calibration curve (figure 1C). As in the recalibration procedure, we estimate implied thresholds $g_a(t_i)$ at a fixed therapeutic threshold t_i by fitting a calibration curve g_a for each group *a* (figure 1C). Then, for each threshold *i* we obtain TCE(i,a):

$$TCE(i, a) = t_i - g_a$$

A negative TCE indicates risk underestimation, since the threshold applied to the risk score is lower than the observed incidence of the outcome at that predicted risk level. Similarly, a positive TCE indicates risk overestimation.

To understand the tradeoff between TCE, FPR and FNR, we calculate intergroup SD (IGSD) between the four group-specific values of the three metrics. For a threshold *i*, metric *M* and *A* distinct groups, IGSD_{Mi} is defined as

$$ISGD(M, t_i) = \sqrt{\frac{\sum_{a=1}^{A}(M_{ia} - \mu_{M_i})^2}{A}}, \text{where}\, \mu_{M_i} = \frac{\sum_{a=1}^{A} M_{ia}}{A}$$

IGSD captures the degree of performance disparity between groups; high IGSD in FPR and FNR corresponds to an EO disparity, and high IGSD in TCE corresponds to a treatment rule disparity.

For each of the four subgroups, and overall population, we report calibration and discrimination metrics at both the aggregate (absolute calibration error, ACE[22] and area under the receiver operating characteristic, AUROC) and the threshold level (TCE, FPR and FNR) at t_1 and t_2, for the UC model, the group-recalibrated model (rUC) and the best-performing EO model, as well as the original PCEs[9] (PCE) and revised PCEs[11] (rPCE). We draw 1000 bootstrap samples from the test set, stratified by group and outcomes, to derive point estimates and 95% CIs for each metric. The 95% CIs are defined as the 2.5% to 97.5% percentiles of the distribution obtained via pooling over both the bootstrap samples and the 10 model replicates derived from the training procedure. We also report IGSD between the four group-specific median values in TCE, FPR and FNR at both thresholds. All metrics are computed over the uncensored population and adjusted for censoring using IPCW.

RESULTS

We describe the study population and present performance of the models. We report the TCE, FPR and FNR in figure 3, and IGSD of those three metrics in figure 4. We present results for EO3 in figure 3, as it was the only equalised odds model that achieved a reduction of IGSD (FPR) while keeping a low IGSD (FNR) at both thresholds. Results for the remaining EO models are included in online supplemental file B.

Study population

Overall, 25 619 individuals met the inclusion criteria, of whom 80% (N=20495) were assigned to the training set, and 10% (N=2562) to each the recalibration and test sets. Table 1 summarises the mean age, ASCVD event incidence and frequency of censoring across the six data sets and four demographic groups. A cohort construction flowchart is included in online supplemental figure B1.

Model performance

The UC model achieved an overall AUROC of 0.827, (95% CI=(0.800 to 0.853)), comparing favourably with PCE (0.808 (0.779 to 0.835)) and rPCE (0.804 (0.777 to 0.831)), while maintaining differences of AUROC between groups (figure 3A). While UC had a slightly higher overall ACE (0.011 (0.006 to 0.023)) than rPCE (0.005 (0.001 to 0.015)), as well as a slightly higher local miscalibration at t_1 (TCE(t_1) 0.012 (0.006 to 0.019) versus 0.000 (−0.004 to 0.005)), IGSD(TCE, t_1) and IGSD(TCE, t_2) both reduced under UC (from 0.018 to 0.004, and 0.053 to 0.016, respectively) (figure 4).

The group recalibration procedure (rUC) reduced the magnitude of TCE(t_1) overall (−0.001 (−0.007 to 0.006)), and for each group, relative to UC (0.012 (0.006

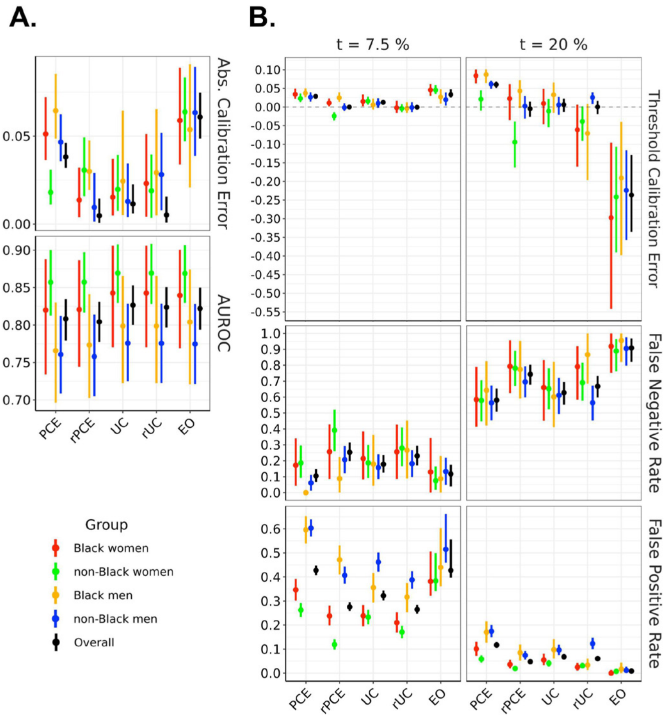


Figure 3 Model performance across evaluation metrics, stratified by demographic group, evaluated on the test set. The left panel showsAUROC and absolute calibration error. The right panel shows false negative rates, false positive rates and threshold calibration error at two therapeutic thresholds (7.5% and 20%). EO, equalised odds; PCEs, original pooled cohort equations; rPCE, revised PCEs; rUC, recalibrated model; UC, unconstrained model.

to 0.019)) (figure 3A). While recalibration improved $TCE(t_2)$ overall (0.00 (−0.019 to 0.016) vs 0.006 (−0.013 to 0.023)), it increased the magnitude of miscalibration of individual groups—for instance, shifting $TCE(t_2)$ from 0.033 (−0.016 to 0.066) to −0.071 (−0.196 to 0.008) for Black men, and increasing $IGSD(TCE, t_2)$ to 0.038.

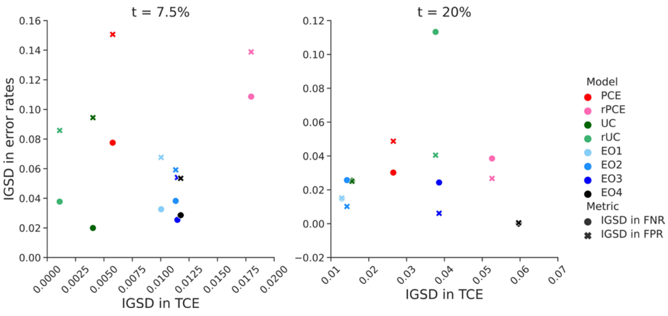


Figure 4 Relationship between intergroup variability in threshold calibration rate (TCE) and error rates. The figure shows the relationship between intergroup SD (IGSD) of threshold calibration error (on the x-axis) and IGSD of false negative rate (FNR, circles) and false positive rate (FPR, crosses) across the models: EO1–4, equalised odds with increasing values of λ. The EO3 corresponds to the EO model discussed in the Results section. In the models we trained, IGSD of TCE scales inversely with the IGSD of FNR and FPR. PCE, original pooled cohort equations; rPCE, revised PCEs; rUC, recalibrated model; UC, unconstrained model.

We also observe that, while $TCE(t_1)$ and $IGSD(TCE, t_1)$ improved for rUC, $IGSD(FNR, t_1)$ worsened, increasing from 0.020 to 0.038, as did $IGSD(FPR, t_2)$ and $IGSD(FNR, t_2)$ (figure 4). Additionally, at each threshold, for all models, we observe a relationship between TCE, FPR and FNR: increased TCE (overestimation) leads to higher FPR and lower FNR, and decreased TCE (underestimation)—to lower FPR an higher FNR (figure 3B).

The EO procedure generated models with FPR and FNR which approached similar values across groups at both t_1 or t_2—bringing $IGSD(FPR, t_1)$ to 0.054 from 0.094 (figure 4) while maintaining almost identical AUROC to UC (0.822 (0.794, 0.8)) (figure 3A). However, it did so by trading off error rates in opposite directions at the two thresholds, as described above (figure 3B). It also increased the magnitude of TCE at both thresholds (from 0.012 (0.006 to 0.019) to 0.033 (0.025 to 0.047) at t_1 and from 0.006 (−0.013 to 0.023) to −0.236 (−0.335 to −0.129) at t_2), and increased $IGSD(TCE, t_1)$ to 0.011 (from 0.004) and $IGSD(TCE, t_2)$ to 0.039 (from 0.016), implying that the scores generated by the EO model did not closely correspond to their calibrated values.

DISCUSSION

We identified local calibration of 10-year ASCVD risk prediction models at guideline-recommended thresholds as necessary for fair shared decision-making about statin treatment between patients and physicians. We find that the rPCEs[11] differ in local calibration between groups—making guideline-compatibility of rPCE inconsistent across groups. We note that global measurements of calibration, used previously to evaluate the PCEs,[11 14] did not capture this difference, illustrating the importance of local calibration evaluation.

Recalibrating the model separately for each group increased compatibility with guidelines at low levels of risk, while increasing intergroup differences in error rates. Conversely, estimators learnt with an EO constraint would not be concordant with existing guidelines as a result of induced miscalibration. Thus, absent a contextual analysis, fairness approaches that focus on error rates can produce misleading results.

In our experiments, group-recalibration did not improve calibration at t=20%. This may be due to the small sample size of the recalibration set, as well as of individuals predicted to be at high risk. This suggests that group-recalibration may not always be desirable, especially if local calibration of the UC model is deemed acceptable. However, improvement in local calibration observed at t=7.5% may be more relevant than calibration at higher risk levels for informing statin initiation decisions, since benefits of treatment are clearer at higher-risk levels.

Several design choices may have impacted the results, including the use of a single model with race and sex as variables in the UC and EO models, the use of a logistic regression as a recalibration method, and the use of an in-processing method that focused on particular decision

Table 1 Cohort characteristics for patients who met inclusion criteria

	N	Age	ASCVD event incidence*	% censored	N	Age	ASCVD event incidence*	% censored
Study	**Black women**				**Black men**			
ARIC	1812	53.2	5.70%	6.51	1216	53.8	9.61%	10.03
CARDIA	232	43.0	4.42%	8.19	153	42.7	1.63%	14.38
CHS	304	70.7	22.52%	15.46	181	70.5	30.89%	27.62
JHS	1310	51.4	2.77%	14.96	751	51.1	4.47%	14.11
MESA	768	60.3	5.18%	9.64	630	60.9	7.19%	13.17
All	4426	54.6	5.69%	10.26	2931	55.1	8.15%	13.07
	Non-Black women				**Non-Black men**			
ARIC	4815	53.9	2.54%	3.30	4383	54.5	7.17%	4.86
CARDIA	289	42.7	0.39%	6.23	333	42.5	0.90%	6.91
CHS	1848	70.7	20.18%	15.58	1169	71.0	32.00%	17.45
FHS OS	828	46.4	2.61%	1.81	856	47.1	8.67%	3.86
MESA	1913	60.5	3.81%	7.68	1828	60.8	6.67%	10.07
All	9693	57.4	5.95%	6.47	8569	56.9	10.36%	7.67
	All							
All	25 619	56.5	7.54%	8.28				

Data are grouped by sex and race, as well as data set. Each group of patients is described by four values: total number of individuals, mean age, censoring-adjusted incidence of ASCVD events within 10 years of the initial examination and fraction of censored individuals.

*ASCVD event incidence was calculated by weighing the number of positive outcome and negative outcome uncensored individuals with the sum of their inverse probability of censoring weights.

ARIC, Atherosclerosis Risk in Communities Study; ASCVD, atherosclerotic cardiovascular disease; AUROC, area under the receiver operating characteristic; CARDIA, Coronary Artery Risk Development in Young Adults Study; CHS, Cardiovascular Health Study; FHS OS, Framingham Heart Study Offspring Cohort; FNR, false negative rate; FPR, false positive rate; IPCW, inverse probability of censoring; JHS, Jackson Heart Study; MESA, Multi-Ethnic Study of Atherosclerosis; PCE, Pooled Cohort Equations.

thresholds to impose EO. We anticipate that alternative modelling choices would impact the size of the observed effects, but would likely not change the conclusions, since known statistical tradeoffs exist between EO and calibration.[18 21 22]

Given this analysis, we recommend that developers building models for use with the ACC/AHA guidelines prioritise calibration across a relevant range of thresholds, and report group-stratified evaluation of local calibration alongside metrics of global fit. Before a model is deployed in a new setting, we recommend that it be evaluated on the target population, stratified by relevant groups—and group-recalibrated, if necessary. Knowledge about local miscalibration should also be incorporated into risk calculators to actively inform the physician-patient shared decision-making conversations, but should not replace recalibration efforts, since calibrated predictions are better suited for reasoning about potential consequences of treatment.[10]

Our analysis inherits the assumptions about relative importance of relevant risks and benefits used to derive therapeutic thresholds (online supplemental file C), which often fail to consider the impact of social determinants of health on treatment efficacy and of structural forms of discrimination in generating health disparities.[28] Additionally, our use of self-identified racial categories—which can be understood as proxies for systemic and structural racist factors impacting health—may be inappropriate, potentially exacerbating historical racial biases and disparities in the clinical settings.[29 30] Derivation of new risk prediction models may be necessary for multiethnic populations.[12] Future work should explore decision analysis and modelling choices that incorporate this context.

CONCLUSION

Our analysis is one of the first to consider algorithmic fairness in the context of clinical practice guidelines. It illustrates general principles that can be used to identify contextually relevant fairness evaluations of models used in clinical settings in the presence of clinical guidelines. Such analysis should include careful consideration of the interplay between model properties, model-guided treatment policy, as well as the potential harms and benefits of treatment, for each relevant subgroup. At the same time, we note that striving for model fairness is unlikely to be sufficient in addressing health inequities, especially when their sources lay upstream of the model-guided intervention, as is the case of structural racism.[28] We encourage future work to situate fairness analyses in this broader context.

Acknowledgements We thank Scotty Fleming, Alison Callahan, Sanjana Srivastava, Sherri Rose, Jonathan Lu, Ethan Steinberg, Yizhe Xu, Jennifer Wilson, Nikolaos Ignatiadis, Steve Yadlowsky and Julian Genkins for insightful comments

and discussion. Approval for this non-human subjects research study is provided by the Stanford Institutional Review Board, protocol IRB-46829.

Contributors AF and SRP conceived of the presented idea, and developed it with support from BP and NS. AF cleaned and preprocessed the data, trained predictive models, generated and analysed model evaluations. SRP provided code for training and evaluation and analysed model evaluations. AF drafted the manuscript, and revised it with SRP in consultation with BP and NS. NS supervised the project. AF is the guarantor of the project.

Funding This study was funded by Stanford University School of Medicine (Program for AI in Healthcare); National Heart, Lung, and Blood Institute (R01 HL144555).

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 may be obtained from a third party and are not publicly available. The six data sets used: ARIC (Atherosclerosis Risk in Communities Study), CARDIA (Coronary Artery Risk Development in Young Adults Study), CHS (Cardiovascular Health Study, 1989–1999), FHS OS (Framingham Heart Study Offspring Cohort), MESA (Multi-Ethnic Study of Atherosclerosis) and JHS (Jackson Heart Study), are available on BioLINCC.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

ORCID iDs
Agata Foryciarz http://orcid.org/0000-0002-8968-5805
Nigam Shah http://orcid.org/0000-0001-9385-7158

REFERENCES

1 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.
2 Barocas S, Hardt M, Narayanan A. *Fairness and machine learning*. fairmlbook.org, 2019.
3 Hardt M, Price E, Srebro N. Equality of opportunity in supervised learning. *Proceedings of the 30th International Conference on Neural Information Processing Systems*, 2016. p.:3315–23.
4 Arnett DK, Blumenthal RS, Albert MA, *et al*. 2019 ACC/AHA Guideline on the Primary Prevention of Cardiovascular Disease: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. *J Am Coll Cardiol* 2019;74:e177–232.
5 Lloyd-Jones DM, Braun LT, Ndumele CE, *et al*. Use of risk assessment tools to guide decision-making in the primary prevention of atherosclerotic cardiovascular disease: a special report from the American heart association and American College of cardiology. *Circulation* 2019;139:e1162–77.
6 Wilson PWF, Polonsky TS, Miedema MD, *et al*. Systematic review for the 2018 AHA/ACC/AACVPR/AAPA/ABC/ACPM/ADA/AGS/APhA/ASPC/NLA/PCNA guideline on the management of blood cholesterol: a report of the American College of Cardiology/American heart association Task force on clinical practice guidelines. *J Am Coll Cardiol* 2019;73:3210–27.
7 Pauker SG, Kassirer JP. Therapeutic decision making: a cost-benefit analysis. *N Engl J Med* 1975;293:229–34.
8 Goff DC, Lloyd-Jones DM, Bennett G, *et al*. 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. *J Am Coll Cardiol* 2014;63:2935–59.
9 Stone NJ, Robinson JG, Lichtenstein AH, *et al*. 2013 ACC/AHA guideline on the treatment of blood cholesterol to reduce atherosclerotic cardiovascular risk in adults: a report of the American College of Cardiology/American heart association Task force on practice guidelines. *J Am Coll Cardiol* 2014;63:2889–934.
10 Cook NR, Ridker PM. Calibration of the pooled cohort equations for atherosclerotic cardiovascular disease: an update. *Ann Intern Med* 2016;165:786–94.
11 Yadlowsky S, Hayward RA, Sussman JB, *et al*. Clinical implications of revised pooled cohort equations for estimating atherosclerotic cardiovascular disease risk. *Ann Intern Med* 2018;169:20–9.
12 Rodriguez F, Chung S, Blum MR, *et al*. Atherosclerotic cardiovascular disease risk prediction in disaggregated Asian and Hispanic subgroups using electronic health records. *J Am Heart Assoc* 2019;8:e011874.
13 et alPfohl SR, Rodriguez F, Marafino B. Creating fair models of atherosclerotic cardiovascular disease. *Proceedings of the 2019 AAAI/ACM Conference on AI*, Ethics, and Society, 2019:271–8.
14 Barda N, Yona G, Rothblum GN, *et al*. Addressing bias in prediction models by improving subpopulation calibration. *J Am Med Inform Assoc* 2021;28:549–58.
15 Pandya A, Sy S, Cho S, *et al*. Cost-Effectiveness of 10-year risk thresholds for initiation of statin therapy for primary prevention of cardiovascular disease. *JAMA* 2015;314:142–50.
16 Yebyo HG, Aschmann HE, Puhan MA. Finding the balance between benefits and harms when using statins for primary prevention of cardiovascular disease: a modeling study. *Ann Intern Med* 2019;170:1–10.
17 Bakalar C, Barreto R, Bergman S. Fairness On The Ground: Applying Algorithmic Fairness Approaches to Production Systems. arXiv [cs.LG], 2021. Available: http://arxiv.org/abs/2103.06172
18 Simoiu C, Corbett-Davies S, Goel S. The problem of infra-marginality in outcome tests for discrimination. *Ann Appl Stat* 2017;11:1193–216.
19 Tripathi S, Fritz BA, Abdelhack M. (Un)fairness in Post-operative Complication Prediction Models. arXiv [cs.LG], 2020. Available: http://arxiv.org/abs/2011.02036
20 Hastings JS, Howison M, Inman SE. Predicting high-risk opioid prescriptions before they are given. *Proc Natl Acad Sci U S A* 2020;117:1917–23.
21 Corbett-Davies S, Goel S. The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. arXiv [cs.CY], 2018. Available: http://arxiv.org/abs/1808.00023
22 Pfohl SR, Foryciarz A, Shah NH. An empirical characterization of fair machine learning for clinical risk prediction. *J Biomed Inform* 2021;113:103621.
23 van der Laan MJ, Robins JM. *Unified methods for censored longitudinal data and causality*. 2003rd ed. New York, NY: Springer, 2003.
24 Cotter A, Jiang H, Gupta MR. Optimization with Non-Differentiable constraints with applications to Fairness, recall, Churn, and other goals. *J Mach Learn Res* 2019;20:1–59.
25 Woodworth B, Gunasekar S, Ohannessian MI. Learning Non-Discriminatory predictors. *Proceedings of the 2017 Conference on Learning Theory*, 2017. p.:1920–53.
26 Pedregosa F, Varoquaux G, Gramfort A. Scikit-learn: machine learning in python. *The Journal of Machine Learning research* 2011;12:2825–30.
27 Paszke A, Gross S, Massa F. PyTorch: an imperative style, high-performance deep learning library. In: *Advances in neural information processing systems*. Curran Associates, Inc, 2019.
28 Churchwell K, Elkind MSV, Benjamin RM, *et al*. Call to action: structural racism as a fundamental driver of health disparities: a presidential Advisory from the American heart association. *Circulation* 2020;142:e454–68.
29 Vyas DA, Eisenstein LG, Jones DS. Hidden in Plain Sight - Reconsidering the Use of Race Correction in Clinical Algorithms. *N Engl J Med* 2020;383:874–82.
30 Hicken MT, Kravitz-Wirtz N, Durkee M, *et al*. Racial inequalities in health: framing future research. *Soc Sci Med* 2018;199:11–18.

BMJ Health & Care Informatics

Investigating for bias in healthcare algorithms: a sex-stratified analysis of supervised machine learning models in liver disease prediction

Isabel Straw , Honghan Wu

To cite: Straw I, Wu H. Investigating for bias in healthcare algorithms: a sex-stratified analysis of supervised machine learning models in liver disease prediction. *BMJ Health Care Inform* 2022;**29**:e100457. doi:10.1136/bmjhci-2021-100457

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bmjhci-2021-100457).

Received 31 July 2021
Accepted 06 April 2022

Institute of Health Informatics, University College London, London, UK

Correspondence to
Dr Isabel Straw;
isabelstraw@doctors.org.uk

ABSTRACT

Objectives The Indian Liver Patient Dataset (ILPD) is used extensively to create algorithms that predict liver disease. Given the existing research describing demographic inequities in liver disease diagnosis and management, these algorithms require scrutiny for potential biases. We address this overlooked issue by investigating ILPD models for sex bias.

Methods Following our literature review of ILPD papers, the models reported in existing studies are recreated and then interrogated for bias. We define four experiments, training on sex-unbalanced/balanced data, with and without feature selection. We build random forests (RFs), support vector machines (SVMs), Gaussian Naïve Bayes and logistic regression (LR) classifiers, running experiments 100 times, reporting average results with SD.

Results We reproduce published models achieving accuracies of >70% (LR 71.31% (2.37 SD) – SVM 79.40% (2.50 SD)) and demonstrate a previously unobserved performance disparity. Across all classifiers females suffer from a higher false negative rate (FNR). Presently, RF and LR classifiers are reported as the most effective models, yet in our experiments they demonstrate the greatest FNR disparity (RF; −21.02%; LR; −24.07%).

Discussion We demonstrate a sex disparity that exists in published ILPD classifiers. In practice, the higher FNR for females would manifest as increased rates of missed diagnosis for female patients and a consequent lack of appropriate care. Our study demonstrates that evaluating biases in the initial stages of machine learning can provide insights into inequalities in current clinical practice, reveal pathophysiological differences between the male and females, and can mitigate the digitisation of inequalities into algorithmic systems.

Conclusion Our findings are important to medical data scientists,clinicians and policy-makersinvolved in the implementationmedical artificial intelligence systems. Anawareness of the potential biases of these systemsmis essential in preventing the digital exacerbation ofhealthcare inequalities.

Summary

What is already known on this topic

⇒ Machine learning models that leverage biochemical data for modelling patient trajectories are rapidly increasing, yet these algorithms are rarely scrutinised for demographic bias or their impact on health inequalities.

What this study adds

⇒ Our study demonstrates a previously unobserved sex disparity in model performance for algorithms built from a commonly used liver disease dataset. We highlight how biochemical algorithms may reinforce and exacerbate existing healthcare inequalities.

How this study might affect research, practice or policy

⇒ Bias in biochemical algorithms is an overlooked issue. In clinical practice, the higher rate of false negatives for female patients would manifest as an increased rate of missed diagnosis for female patients and a consequent lack of appropriate care.

⇒ Furthermore, sex differences in biochemical feature importance reinforces existing research that suggests unisex biochemical thresholds may disadvantage female patients in current practice. These findings are important to medical data scientists, clinicians and policy-makers involved in the implementation medical artificial intelligence systems. An awareness of the potential biases of these systems is essential in preventing the digital exacerbation of healthcare inequalities

BACKGROUND

Liver cirrhosis accounts for 1.8% of deaths in Europe, a number which has grown significantly over the past decade as rates of alcohol consumption, chronic hepatitis infections and obesity-related liver disease have increased.[1] Yet, liver disease does not affect all populations equally. Recent research has demonstrated sex differences in the prevalence, diagnosis and management of various hepatic illnesses.[2–5] A key determinant of patient outcomes from liver disease is the early detection of pathology, yet when it comes to diagnosis and referral, female patients appear to be at a significant disadvantage.[2–5]

In alcohol related liver disease, Vatsalya *et al* report that women are less likely to be suspected of alcohol abuse, diagnosed and often experience more severe disease with worse outcomes.[2 3] Sex differences in diagnosis are compounded by inequalities in the liver disease management. Mathur *et al* report disparities in access to liver transplantation that result in females having markedly lower transplant rates than their male counterparts.[4] The problem extends beyond hepatology. In 2021, the UK parliamentary report on the gender health gap highlighted that the UK has the largest female health gap in the G20 and the 12th largest globally.[5] The exclusion of females from research trials (extending to animal research), the neglect of female bodies throughout medical pedagogy and the unconscious biases of practitioners are a few of the intersecting factors that result in worse health outcomes for female patients.[6–10]

Liver function tests are integral to patient diagnosis and monitoring. These 'biochemical markers' include proteins made by the liver (eg, albumin), and enzymes required for metabolism (eg, aspartate aminotransferase (AST)). Bias research has illustrated that biochemical markers are not equally effective for all patient groups.[3 7 10–12] Suthahar *et al* describe how sex differences in biomarker thresholds affect objectivity in management, as what is considered 'normal' in one sex, may not be so in the other.[12] Grimm *et al* investigate the relationship between albumin and mortality, reporting that albumin offers a higher predictive power for males compared with females.[11] Furthermore, Vatsalya *et al* and Stepien *et al* describe sex differences in biochemical cut offs, highlighting that the milder expression of liver injury for females may result in female disease going undetected.[3 13] Such disparities in the predictive potential of clinical biomarkers have the potential to exacerbate healthcare inequalities.[6 7 10 12]

The rise in healthcare artificial intelligence (AI) has resulted in the increasing use of large clinical datasets for machine learning (ML).[14] ML classifiers that use biochemical markers to model patient trajectories have consistently outperformed traditional statistical models.[14] However, despite the promise of ML tools, the presence of demographic biases in AI algorithms has indicated that historical harms may materialise in digital systems and worsen population inequalities.[7 15–17] The development of predictive models from biomarkers is one area in which medical ML models are at risk of encoding the errors of current practice. In our paper we explore for this possibility in liver disease prediction by examining models built from a commonly cited dataset: The Indian Liver Patient Dataset (ILPD).

The ILPD is a widely used open-source dataset that provides the biochemical markers of a sample of patients, some of whom have liver disease.[18–22] BanuPriya and Tamilselvi provide an overview of classification models built from this dataset, since which time further models have been published from both academics and major industry.[18 19 21] Authors consistently report accuracies of >70% for identifying liver patients, with logistic regression (LR) models and random forests (RFs) giving the best results. Jin *et al*[23] demonstrate accuracies of 72.7% with LR models, similarly Adil *et al* achieve 74% accuracy with their LR model, outperforming artificial neural networks and support vector machines (SVMs).[24] A recent study from Intel reproduces these models and performs additional feature selection giving model accuracies of 74.6% (RF) and 71.2% (SVM).[19]

Predictive ML models may benefit patient care if they can diagnose liver disease at an earlier stage.[25] Yet, despite the existing literature that describes biases in clinical medicine, biochemical tests and algorithmic performance, none of the ML studies on the ILPD focus on sex disparities in model performance.[4 7 8 10–12 16 17] We seek to address this gap in the research by investigating the ILPD dataset and its respective models for sex bias.[18–20]

METHODOLOGY

The ILPD was originally collected from India and consists of 583 patient records, of which 416 have liver disease. We imported the ILPD from the UCI repository (full codebook available in online supplemental material C).[19 22]

Data exploration and initial analysis

Data exploration is the primary stage of the ML process and involves file importation, formatting, descriptive statistics and configuring datatypes. Online supplemental table 1 gives the variables included in our dataset and their initial datatypes.

Feature exploration

Online supplemental table 2 presents the sex-stratified feature importance ranked by Pearson's correlation coefficient. For females, the enzymes ALT and AST are ranked fourth and fifth, whereas for males they are ranked seventh and eighth. Further, albumin and A/G ratio are ranked higher for male patients compared with female patients. These subtle differences in feature importance may reflect underlying sex differences in hepatic pathophysiology and biomarker expression.[3 4 26] Further, online supplemental table 2 demonstrates that the mean IQR across all biomarkers is less for females, suggesting that these biomarkers may have less of a predictive power for female patients overall (mean IQR; female 0.145, male 0.175).

Data preprocessing

Data preparation steps reflected existing studies.[19 20] Mean imputation was used to address missing values, gender was mapped to a 0/1 numerical datatype, normalisation was performed using minimum-maximum scaler function and the target variable was recoded to binary variable, such that 1 represents diseased patients (n=416).

Addressing class imbalance

The original dataset demonstrated significant class imbalance (167 healthy vs 416) diseased patients) and sex imbalance (142 females vs 441 males). Similarly to existing models, we implement the imblearn SMOTE()

Table 1 Summary counts of classes in the Indian liver patient dataset dataset, including counts after the dataset is balanced

	Target (disease=1)	Dataset 1 (original)	Total counts for sexes	Dataset 2 (oversampled minority class)	Total counts for sexes	Dataset 3 (sex balanced, oversampled females)	Total counts for sexes
Female	0	50	142	145	237	408	595
	1	92		92		187	
Male	0	117	441	271	595	271	595
	1	324		324		324	
Total			583		832		1190

package to address these imbalances; oversampling both the minority class and under-represented females as detailed in table 1.[19] The sex-unbalanced dataset is retained to compare the impact of female representation in the training data on sex disparities in performance.

Model development and implementation

Gulia and Praveen Rani review the classification algorithms that have been built from the ILPD, including RFs and SVMs.[20] A more recent review from BanuPriya and Tamilselvi describe the accuracies of additional models including Bayesian Networks, which is further built on by the work of Aswathy who evaluates the performance of LR models on the ILPD.[18 19] We replicate the methods of these studies, reproducing RF, SVM, Gaussian Naïve Bayes (GNB) and LR classifiers. We implement these models across four experiments, in which we evaluate the overall and sex-stratified performance of the classifiers.

Experiment 1: models trained on unbalanced dataset, without feature selection

Initially, we reproduce existing studies, building a predictive algorithm on the full unbalanced dataset to predict liver disease. Data were divided into test and training subsets (30%/70%), hyperparameters were tuned using GridSearchCV(), the model was trained on the mixed-sex data and results were stratified by sex to give the evaluation metrics for males/females separately. We do this 100 times (building, training and testing separate models) and report average results with SD over the 100 runs. This is done for all four classifiers resulting in four results tables (online supplemental material B Spreadsheets, 'Experiment 3.1.1—RF'—'Experiment 3.1.1 GNB').

Experiment 2: models trained on sex-balanced dataset, without feature selection

The methodology of experiment 1 is repeated using the sex-balanced dataset defined in Table 1 . We ensure sex balance in the training data by taking random subsets from the male and females separately, which are appended together to form the full sex-balanced training data for each individual experiment (online supplemental file 3 Spreadsheets, 'Experiment 3.1.2—RF'—'Experiment 3.1.2 GNB').

Experiment 3: models trained on unbalanced dataset, with feature selection

In experiment 3, we perform feature selection based on the unbalanced dataset, in experiment 4, we perform feature selection on the sex-balanced dataset. Feature selection is performed using Recursive Feature Elimination (RFE) sklearn package, which returns the top five ranked features (online supplemental material B Spreadsheets, 'Experiment 3.1.3—RF'—'Experiment 3.1.3 GNB').

Experiment 4: models trained on balanced dataset, with feature selection

Lastly, models and feature selection are fitted to the sex-balanced dataset. Our aim was to investigate whether feature selection would differ once the representation of females was addressed, and whether this would influence any performance disparities.

Model evaluation

Evaluation metrics are reported for all patients and separately for the sexes (equations 1–3). We examine the mean difference between the male and females for each evaluation metric to demonstrate any disparities (equation 4). Two-sample paired t-tests are run on the series of 100 experiments for the male and female patients to assess whether the mean difference between sexes, for each of the evaluation metrics, is statistically significant ($p<0.05$).

Equation 1: accuracy evaluation metric

Accuracy gives the proportion of correct predictions produced by a model.

$$\text{Accuracy} = \frac{\text{True positives+True Negatives}}{\text{True positives+True Negatives+False Positives+False Negatives}}$$

Equation 2: F-score evaluation metric, precision and recall

The F-score is the average of precision and recall, with a value of 1 being a perfect score.

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}}$$

$$\text{F Score} = \frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision+Recall}}$$

Equation 3: performance error rates

The following error rates are used throughout our evaluation.[21]

- True positive: Predicted yes and they do have disease.
- True negative: Predicted no and they do not have disease.
- False positive: Predicted yes, but they do not have disease.
- False negative: Predicted no, but they actually do have disease.

$$\text{True Negative Rate }(\text{TNR}) = \frac{\text{TN}}{\text{TN+FP}}$$

$$\text{True Positive Rate }(\text{TPR}) = \frac{\text{TP}}{\text{TP+FN}}$$

Equation 4: sex performance disparity

$$\text{Sex performance disparity} = \text{Male evaluation metric }(\text{mean}) - \text{Female evaluation metric }(\text{mean})$$

RESULTS

We ran 16 experiments: experiments 1–4, with each of the four classifiers. The detailed results tables with the 100 experiment runs are provided in the spreadsheet files in online supplemental material B. In online supplemental material A 'Tables in Text', we provide summary in several condensed tables, which give the average evaluation metrics and the statistical significance of any male-female differences.

Results for experiment 1

Online supplemental table 3 demonstrates that our four models reflect the existing literature, achieving accuracies >70% (71.31% (2.37 SD) LR – 79.40% (2.50 SD) SVM). Table 2 details the disparities for each evaluation metric, from which we observe a statistically significant sex disparity in Accuracy for all classifiers, with mixed results regarding the direction of the disparity (performance disparity −2.98 SVM to 2.96% RF, $p<0.05$). In the case of the ROC_AUC score, we observe a significant disparity that negatively impacts females for the RF (6.80%, $p<0.05$), LR (2.93%, $p<0.05$) and GNB (5.53%, $p<0.05$) classifiers.

The accuracy and ROC_AUC disparities fluctuate depending on the balance between the different error rates, however, on examining the error rates individually, we see a consistency in error trends for each sex. Across all classifiers females suffer from a higher false negative rate (FNR), while males suffer from a higher false positive rate. The disparity demonstrates a consistently higher recall for males, with females experience a lower recall and correspondingly higher FNR disparity, −2.58% to −24.07%, table 2)

Results for experiment 2

In experiment 2, we trained on sex-balanced data, improving overall accuracy across all four classifiers (RF 81.66% (2.33 SD) vs 78.17 (2.36 SD); LR 74.53% (1.96 SD) vs 71.31% (2.37 SD); SVM 83.30% (1.75 SD) vs 79.40% (2.50 SD); GNB 74.75% (1.9 SD) vs 71.53% (2.61 SD)—online supplemental table 4). We now see a consistent accuracy disparity that benefits females across all four classifiers (−11.47% to −6.17%, $p<0.05$–table 3). Disparities in the ROC_AUC scores are less consistent (LR unbalanced ROC disparity 2.93%, LR balanced ROC disparity 4.79%; GNB unbalanced ROC disparity 5.53%, GNB balanced disparity 5.45%).

Table 2 Experiment 3.1.1—unbalanced training data without feature selection, sex performance disparities

	Random forest classifier		Logistic regression classifier		Support vector machine		Gaussian Naïve Bayes	
Mean difference averaged over n=100	**Sex performance disparities (%)**	**t-test p value**	**Sex performance disparities (%)**	**t-test p value**	**Sex performance disparities (%)**	**t-test p value**	**Sex performance disparities (%)**	**t-test p Value**
Accuracy	2.96	0.00	−2.85	0.01	−2.98	0.02	−2.72	0.02
FScore	15.63	0.00	15.86	0.00	4.14	0.00	16.19	0.00
ROC_AUC*	6.80	0.00	2.93	0.00	−2.41	0.08	5.53	0.00
Precision	5.25	0.00	−4.87	0.00	3.41	0.00	−3.13	0.05
Recall	21.02	0.00	24.07	0.00	2.58	0.04	19.31	0.00
False negative rate	−21.02	0.00	−24.07	0.00	−2.58	0.08	−19.31	0.00
True negative rate	−7.42	0.00	−18.20	0.00	−7.40	0.00	−8.24	0.00
False positive rate	7.42	0.00	18.20	0.00	7.40	0.00	8.24	0.00
True positive rate	21.02	0.00	24.07	0.00	2.58	0.04	19.31	0.00

*ROC AUC score is a measure of the separation between classes in a binary classifier, derived from the area under the ROC curve.

Table 3 Experiment 3.1.2—balanced training data without feature selection, sex performance disparities

Mean difference averaged over n=100	Random forest classifier		Logistic regression classifier		Support vector machine		Gaussian Naïve Bayes	
	Sex performance disparities (%)	t-test p value	Sex performance disparities (%)	t-test p value	Sex performance disparities (%)	t-test p value	Sex performance disparities (%)	t-test p value
Accuracy	−6.17	0.00	−6.36	0.00	−11.47	0.00	−7.43	0.00
FScore	7.69	0.00	20.17	0.00	−3.40	0.00	16.65	0.00
ROC_AUC	0.60	0.13	4.79	0.00	−9.06	0.00	5.45	0.00
Precision	−0.94	0.88	−4.75	0.00	−2.32	0.14	0.24	0.37
Recall	12.88	0.00	29.22	0.00	−4.64	0.00	19.82	0.00
False negative rate	−12.88	0.00	−29.22	0.00	4.64	0.00	−19.82	0.00
True negative rate	−11.69	0.00	−19.65	0.00	−13.49	0.00	−8.93	0.00
False positive rate	11.69	0.00	19.65	0.00	13.49	0.00	8.93	0.00
True positive rate	12.88	0.00	29.22	0.00	−4.64	0.00	19.82	0.00

Online supplemental table 5 presents a comparison of the evaluation metrics with/without balancing of the training data. In one case, we observe an improvement in performance for all patients. When trained on the balanced dataset, the LR accuracy improves overall (74.53% (1.96 SD) vs 71.31% (2.37 SD)), for females (77.71% (2.42 SD) vs 73.33% (3.95 SD)) and for males (71.35% (3.22 SD) vs 70.49% (2.74 SD)).

Results for experiment 3

We did not see an improvement in overall performance or a reduction in disparities with RFE. A significant ROC_AUC disparity is apparent across all four classifiers (3.60%–6.61%, p<0.05) that negatively impacts females. We see the same error rate findings as earlier, with a higher FNR for females (FNR Disparity −18.21 to −21.24%, p<0.05, table 4 and online supplemental table 6).

Results for experiment 4

Experiment 4 gives mixed results. The accuracy disparity benefits females across all classifiers (−4.64% to −6.80%, p<0.05), whereas the ROC_AUC disparity demonstrates a benefit for males in three out of four classifiers (−0.05% to 5.95%, p<0.05, table 5) The results relate to the subtle changes in error rates with each model, however, across

Table 4 Experiment 3.1.3—unbalanced training data with feature selection, sex performance disparities

	Random forest classifier		Logistic regression classifier		Support vector machine		Gaussian Naïve Bayes	
	Sex performance disparities (%)	t-test p value	Sex performance disparities (%)	t-test p value	Sex performance disparities (%)	t-test p value	Sex performance disparities (%)	t-test p value
Accuracy	3.42	0.00	−2.90	0.01	−2.75	0.01	−3.31	0.00
FScore	15.36	0.00	15.79	0.00	16.50	0.00	15.29	0.00
ROC_AUC	6.61	0.00	3.60	0.00	4.90	0.00	4.99	0.00
Precision	9.85	0.00	0.24	0.44	−0.87	0.90	−3.41	0.03
Recall	18.21	0.00	21.24	0.00	20.30	0.00	18.54	0.00
False negative rate	−18.21	0.00	−21.24	0.00	−20.30	0.00	−18.54	0.00
True negative rate	−4.99	0.00	−14.04	0.00	−10.50	0.00	−8.57	0.00
False positive rate	4.99	0.00	14.04	0.00	10.50	0.00	8.57	0.00
True positive rate	18.21	0.00	21.24	0.00	20.30	0.00	18.54	0.00

Table 5 Experiment 3.1.4—balanced training data with feature selection, sex performance disparities

	Random forest classifier		Logistic regression classifier		Support vector machine		Gaussian Naïve Bayes	
	Sex performance disparities (%)	t-test p value	Sex performance disparities (%)	t-test p value	Sex performance disparities (%)	t-test p value	Sex performance disparities (%)	t-test p value
Accuracy	−5.62	0.00	−6.80	0.00	−6.19	0.00	−4.64	0.00
FScore	7.86	0.00	14.39	0.00	16.46	0.00	21.63	0.00
ROC_AUC	−0.05%	0.46	3.57%	0.00	5.95%	0.00	8.17%	0.00
Precision	4.60%	0.00	9.28%	0.00	12.82%	0.00	9.35%	0.00
Recall	9.70%	0.00	15.51%	0.00	15.38%	0.00	22.78%	0.00
False negative rate	−9.70	0.00	−15.51	0.00	−15.38	0.00	−22.78	0.00
True negative rate	−9.79	0.00	−8.37	0.00	−3.47	0.00	−6.44	0.00
False positive rate	9.79	0.00	8.37	0.00	3.47	0.00	6.44	0.00
True positive rate	9.70	0.00	15.51	0.00	15.38	0.00	22.78	0.00

all classifiers the FNR is consistently higher for females (−9.70% to −22.78%, p<0.05 (online supplemental table 7).

Analysis of feature selection

Online supplemental table 8 gives the feature rankings assigned by the RFE model when fitted to unbalanced and balanced data, focusing on RF classifiers. When we address the under-representation of females in the training data, ALP and gender are included as the top two features, while A/G ratio and total bilirubin are removed. This finding may reflect existing research that describes sex differences in biomarker expression. In their analysis gender-specific references intervals for hepatic biomarkers, Li *et al* highlight sex differences in ALP, ALT and GGT, indicating that differing thresholds may be appropriate for diagnosis.[27] Sex differences in biochemical disease profiles may explain why integrating more female patients affects the feature selection in experiment 4.

DISCUSSION

In recent years, research has highlighted that medical biases and female under-representation may significantly contribute to differences in healthcare outcomes; in our paper, we have examined how this phenomena may extend into ML.[6–8 10 28] We present several key findings:

- Model reproduction and demonstration of disparity: We have demonstrated a previously unobserved sex disparity that exists in published ML classifiers based on the ILPD dataset.
- Error disparities: Sex disparities in Accuracy and ROC_AUC fluctuate depending on model and the balance between error rates, however, sex differences in specific error rates are consistent. We observe a consistently lower recall and correspondingly higher FNR for females. Of note, RF and LR classifiers are reported as the most effective on the ILPD dataset, however, these models demonstrate the greatest disparity in the FNR when trained on the original dataset (RF, FNR disparity −21.02% (p<0.05); LR, FNR disparity −24.07%, (p<0.05)). Clinically, this FNR disparity would materialise as an inequality in disease detection that negatively impacts females, with higher instances of missed disease.
- Balanced training: Training on sex-balanced data improved overall performance for all classifiers. In the case of the LR classifier, accuracy improves overall and for the sexes separately, indicating that with the right model selection addressing poor performance for the under-represented group does not need to come at the expense of the majority group.
- Impact of model architecture on disparity: Our experimental outcomes were not consistent across models, indicating that bias mitigation techniques may need to be tailored to model choice.
- Analysis of feature ranking: Our comparison of feature importance reinforces existing clinical research that highlights the sex differences in the role of liver biomarkers.

Implications for data science

Our experiments demonstrated that sex-specific feature selection and addressing under-representation of females may be an important bias mitigation technique when developing ML algorithms in medicine. Furthermore, we illustrate that there is no consistent solution across all classifiers, suggesting techniques need to be tailored to model choice. ML models also present novel opportunities for improving existing practice and addressing health disparities that relate to biochemical discrepancies between the sexes. Given the evolving evidence that critiques the use of 'unisex' biochemical thresholds, ML models that do not rely on these defined thresholds may pose a superior alternative if developed with an awareness of the subtle sex differences in disease manifestation.

Implications for clinical medicine and public health

Classification algorithms are being increasingly used in healthcare settings to assist clinicians in medical diagnosis.[20] Unless these algorithms are evaluated for biases, they may only improve care for a subset of patients and consequently increase healthcare inequalities.[7] By evaluating ML models for demographic biases before they are implemented in digital medicine, we can mitigate the perpetuation of these inequalities into digital systems.

Furthermore, insights from model development can be used to inform current clinical care. Our data exploration of feature correlation demonstrated sex differences in feature importance. Such research can inform practising clinicians on the relevance of different indicators for the patient in front of them, for example, albumin may be more indicative of pathology in males.[11] Lastly, examining disparities in algorithmic performance offers an opportunity to reflect on which patients may be being missed in current practice. Throughout our analysis, we demonstrated a persistently high FNR for females, suggesting that female disease is at risk of being overlooked. Examining the physiological profile of algorithmic false negatives presents an opportunity to better understand which patients are at risk of being misdiagnosed.

It should be noted that the ILPD does not include demographic information on race or ethnicity.[22] Racial biases have been reported in the biochemical tests used across different subspecialties, resulting in worse care for marginalised racial groups.[29 30] A key limitation of our study is that we cannot perform a race stratified analysis. Furthermore, we are unable to evaluate the relevance of other demographic features. An intersectional approach to healthcare inequalities would consider the mediating impact of socioeconomic class, or the compounding impact of gender (as opposed to sex) and sexuality on marginalised patients. Accounting for the complex nature of these intersectional relationships requires more advanced modelling and new bias evaluation techniques.

CONCLUSIONS

The historic absence of women from the healthcare profession and from clinical research resulted in domain knowledge that centres around the male body and neglects female physiological differences. To ensure sex-based inequalities do not manifest in medical AI, an evaluation of demographic performance disparities must be integrated into model development. Evaluating biases in the initial stages of ML can provide insights into inequalities in existing practice, reveal pathophysiological differences between the sexes and can mitigate the digitisation of healthcare inequalities in algorithmic systems.

Twitter Isabel Straw @IsabelStrawMD and Honghan Wu @hhwu

Contributors IS conceived of the presented idea and is the guarantor responsible for overall content. IS developed the theory and performed the computations. HW verified the analytical methods.

Funding This work was supported by UK Research and Innovation (UKRI Grant Reference Number EP/S021612/1).

Competing interests None declared.

Patient consent for publication Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available in a public, open access repository.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

ORCID iD

Isabel Straw http://orcid.org/0000-0003-0003-3550

REFERENCES

1 Blachier M, Leleu H, Peck-Radosavljevic M, *et al*. The burden of liver disease in Europe: a review of available epidemiological data. *J Hepatol* 2013;58:593–608.
2 Morgan MY, Sherlock S. Sex-Related differences among 100 patients with alcoholic liver disease. *Br Med J* 1977;1:939–41.
3 Vatsalya V, Liaquat HB, Ghosh K, *et al*. *A review on the sex differences in organ and system pathology with alcohol* drinking. *Curr Drug Abuse Rev* 2016;9:87–92.
4 Mathur AK, Schaubel DE, Gong Q, *et al*. Sex-based disparities in liver transplant rates in the United States. *Am J Transplant* 2011;11:1435–43.
5 UK Parliament, Women's health outcomes: Is there a gender gap?, House of Lords Library, Editor. 2021, House of Lords. Available: https://lordslibrary.parliament.uk/womens-health-outcomes-is-there-a-gender-gap/
6 Cleghorn E. *Unwell women: misdiagnosis and myth in a man-made world*. New York, NY: Dutton, 2021.
7 Straw I. The automation of bias in medical artificial intelligence (AI): decoding the past to create a better future. *Artif Intell Med* 2020;110:101965.
8 Krieger N, Fee E. Man-made medicine and women's health: the biopolitics of sex/gender and race/ethnicity. *Int J Health Serv* 1994;24:265–83.
9 Hoffmann DE, Tarzian AJ. The girl who cried pain: a bias against women in the treatment of pain. *J Law Med Ethics* 2001;29:13–27.
10 Hamberg K. Gender bias in medicine. *Womens Health* 2008;4:237–43.
11 Grimm G, Haslacher H, Kampitsch T, *et al*. Sex differences in the association between albumin and all-cause and vascular mortality. *Eur J Clin Invest* 2009;39:860–5.
12 Suthahar N, Meems LMG, Ho JE, *et al*. Sex-Related differences in contemporary biomarkers for heart failure: a review. *Eur J Heart Fail* 2020;22:775–88.
13 Stepien M, Fedirko V, Duarte-Salles T, *et al*. Prospective association of liver function biomarkers with development of hepatobiliary cancers. *Cancer Epidemiol* 2016;40:179–87.

14 Sidey-Gibbons JAM, Sidey-Gibbons CJ. Machine learning in medicine: a practical introduction. *BMC Med Res Methodol* 2019;19:1–18.

15 Cirillo D, Catuara-Solarz S, Morey C, *et al*. Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. *NPJ Digit Med* 2020;3:81.

16 O'Neil C. *Weapons of math destruction*. Harlow, England: Penguin Books, 2017.

17 Straw I, Callison-Burch C. Artificial intelligence in mental health and the biases of language based models. *PLoS One* 2020;15:e0240376.

18 M. BanuPriya, Tamilselvi PR. *Performance analysis of liver disease prediction using machine learning algorithms*. 5, 2018.

19 Aswathy C. Liver patient dataset classification using the Intel® distribution for python. Intel, specialized development tools, 2018. Available: https://www.intel.com/content/www/us/en/developer/articles/technical/liver-patient-dataset-classification-using-the-intel-distribution-for-python.html

20 Gulia A, Praveen Rani DRV. Liver patient classification using intelligence techniques. *Int J Comput Sci Inf Technol Res* 2014;5:5110–5.

21 Ramana BV, Boddu RSK. *Performance comparison of classification algorithms on medical datasets*. 2019 IEEE 9th Annual computing and communication workshop and conference (CCWC), 2019: 140–5.

22 Dua D, Graff C. UCI machine learning Repository. Irvine, Ca: University of California, school of information and computer science. ILPD dataset, 2019. Available: https://archive.ics.uci.edu/ml/datasets/ILPD+%28Indian+Liver+Patient+Dataset%29#

23 Jin H, Kim S, Kim J. Decision factors on effective liver patient data prediction. *International Journal of Bio-Science and Bio-Technology* 2014;6:167–78.

24 Adil SH, Ebrahim M, Raza K, *et al*. *Liver patient classification using logistic regression*. 4th International Conference on Computer and Information Sciences (ICCOINS). IEEE, 20182018.

25 Auxilla LA. *Accuracy prediction using machine learning techniques for Indian patient liver disease*. 2nd International Conference on Trends in Electronics and Informatics (ICOEII), 2018: 45–50.

26 Guy J, Peters MG. Liver disease in women: the influence of gender on epidemiology, natural history, and patient outcomes. *Gastroenterol Hepatol* 2013;9:633.

27 Li X, Wang D, Yang C, *et al*. Establishment of age- and gender-specific pediatric reference intervals for liver function tests in healthy Han children. *World J Pediatr* 2018;14:151–9.

28 Vyas DA, Eisenstein LG, Jones DS. Hidden in plain sight — reconsidering the use of race correction in clinical algorithms. *N Engl J Med Overseas Ed* 2020;383:874–82.

29 Eneanya ND, Boulware LE, Tsai J, *et al*. Health inequities and the inappropriate use of race in nephrology. *Nat Rev Nephrol* 2022;18:84–94.

30 Powe NR. Black kidney function matters: use or misuse of race? *JAMA* 2020;324:737–8.

BMJ Health & Care Informatics

Can medical algorithms be fair? Three ethical quandaries and one dilemma

Kristine Bærøe ,[1] Torbjørn Gundersen,[2] Edmund Henden,[2] Kjetil Rommetveit[3]

To cite: Bærøe K, Gundersen T, Henden E, *et al.* Can medical algorithms be fair? Three ethical quandaries and one dilemma. *BMJ Health Care Inform* 2022;**29**:e100445. doi:10.1136/bmjhci-2021-100445

Received 11 July 2021
Accepted 10 December 2021

[1]Department of Global Public Health and Primary Care, University of Bergen, Bergen, Norway
[2]Centre for the Study of Professions, Oslo Metropolitan University, Oslo, Akershus, Norway
[3]Center for the Study of the Sciences and Humanities, University of Bergen, Bergen, Hordaland, Norway

Correspondence to
Dr Kristine Bærøe;
kristine.baroe@uib.no

ABSTRACT

Objective To demonstrate what it takes to reconcile the idea of fairness in medical algorithms and machine learning (ML) with the broader discourse of fairness and health equality in health research.
Method The methodological approach used in this paper is theoretical and ethical analysis.
Result We show that the question of ensuring comprehensive ML fairness is interrelated to three quandaries and one dilemma.
Discussion As fairness in ML depends on a nexus of inherent justice and fairness concerns embedded in health research, a comprehensive conceptualisation is called for to make the notion useful.
Conclusion This paper demonstrates that more analytical work is needed to conceptualise fairness in ML so it adequately reflects the complexity of justice and fairness concerns within the field of health research.

Summary

What is already known?

- Biases in data, modelling and human review impact the fairness of the outcome of machine learning (ML).
- ML fairness in healthcare involves the absence of prejudices and favouritism towards an individual or group based on inherent or acquired characteristics, while fairness in health more broadly understood addresses historical fundamental socioeconomic biases that create health inequality within populations.
- Healthcare systems can fairly mitigate unjust health inequalities by offering equal opportunities for healthy lives.

What does this paper add?

- This paper argues that ML fairness in healthcare depends on equal access to healthcare systems.
- It demonstrates how a full conceptualisation of ML fairness in health is conditioned by a complex nexus of different fairness concerns.
- It calls for a reconceptualisation of ML fairness in health that acknowledges this complexity.

INTRODUCTION

Machine learning (ML) refers to algorithms that improve their performance independent of human designers. Several biases are involved in developing and applying ML, such as data biases (eg, historical and representation biases), modelling/design biases (eg, evaluation and aggregation biases) and human review biases (behavioural and social biases).[1 2] Biases affect the fairness of the ML process's outcome and deployment by wrongly skewing the outcome. 'Fairness' can be understood in different ways, but is usefully defined in the context of ML-based decision making as 'the absence of any prejudice or favouritism towards an individual or a group based on their inherent or acquired characteristics'.[3] Thus, when operationalising fairness into ML systems applied in health, the goal should be to eradicate biases in the processes of data sampling, modelling and human review so that the ML process does not promote health advantages or disadvantages for any individuals or groups based on their inherent or acquired characteristics. Assessing what kind of characteristics are assumed relevant for a fairness approach relies on normative ideas about justice. In healthcare, the World Medical Association's Declaration of Geneva identifies 'age, disease or disability, creed, ethnic origin, gender, nationality, political affiliation, race, sexual orientation, social standing or any other factor' as examples of factors that should not impact the doctors' duty towards their patients.[4] Thus, ML failing to perform adequately to certain patients with such characteristics can be judged unfair.

In parallel, a more comprehensive and ambitious conceptualisation of fairness in health is discussed in the literature addressing how to distribute healthcare justly. Fairness is here understood in terms of how healthcare needs are unequally distributed within and across populations in the first place, which calls for justly allocated healthcare to reduce historically and socially conditioned inequalities. Theoretically, this aim is captured by egalitarian approaches to ensure, for example, equal opportunities[5] or capabilities,[6] or social justice.[7 8] Politically, it is reflected in empirically informed reports on observed health inequalities (eg, WHO's report on closing the gap of inequalities in a generation[9] and in the Sustainable Development Goal of promoting health equality[10]). In clinical settings, work has been carried out to clarify the appropriateness of considering socioeconomic factors to circumvent their adverse impact on patients' ability to benefit from treatment.[11] This work has been translated into a call for revising and clarifying the way 'social standing' requires clinical attention in the World Medical Association's Declaration of Geneva.[12]

Unjust health inequality is influenced by inequality in the socioeconomical, cultural and environmental factors (eg, access to clean water) that shape people's living conditions.[13] Although theories diverge as to what makes the resulting health disparities unfair, there is broad consensus that health inequality associated with socioeconomic determinants of health creates inequity and calls for amendment.[14] For this reason, ML fairness should not only be about avoiding prejudices and favouritism, but also about reducing unfair health inequalities,[15] particularly those associated with socioeconomic health determinants . In line with Rajkomar and colleagues' reasoning,[15] to avoid ML in healthcare contributing to maintaining or reinforcing health inequities, fairness should be operationalised into ML processes by ensuring equal outcome across socioeconomic status, equal performance of models across socioeconomic groups, as well as equal allocation of resources.

Against this backdrop, this paper aims to answer the following question: How can the *narrow* fairness discourse related to ML and absence of prejudice and favouritism, and the *broader* fairness discourse related to unjust health equality be reconciled in a comprehensive conceptualisation of ML fairness that can be operationalised to prevent health inequity from being maintained or reinforced by healthcare systems? A more comprehensive notion of fairness in ML healthcare can be used to articulate commitments of fairness and help structure guidelines and recommendations.[16]

We start the discussion by clarifying the nature of the ML algorithms we focus on and present two distinct versions of 'justice' (substantive and procedural). We then argue that an adequate notion of ML fairness depends on a comprehensive approach to *fair access* to healthcare, which is inherently connected with other fairness challenges calling for practical solutions. Next, we identify and describe three interrelated fairness quandaries and one fairness dilemma related to obtaining ML fairness in health. A meaningful conceptualisation of ML fairness, which can be implemented to avoid inequitable patient outcomes, must reflect this complex, intertangled nexus of fairness concerns.

METHOD

The methodological approach used in this paper is theoretical and ethical analysis.

RESULTS

By applying this method, we identify three ethical quandaries and a dilemma related to ML fairness in healthcare. First, there is what we call 'the unfair data quandary'. Second, there is 'the unfair design quandary'. Third, there is 'the reasonable disagreement quandary'. Finally, there is the dilemma that arises from trade-offs between fairness and accountability. Figure 1 illustrates our approach.

DISCUSSION

ML refers to algorithms that improve their performance based on previous results independently of human designers. An important subset of ML with much promise in medicine is deep learning algorithms, which process inputs (eg, data such as pictures, videos, speech and text) to provide output such as identified patterns,

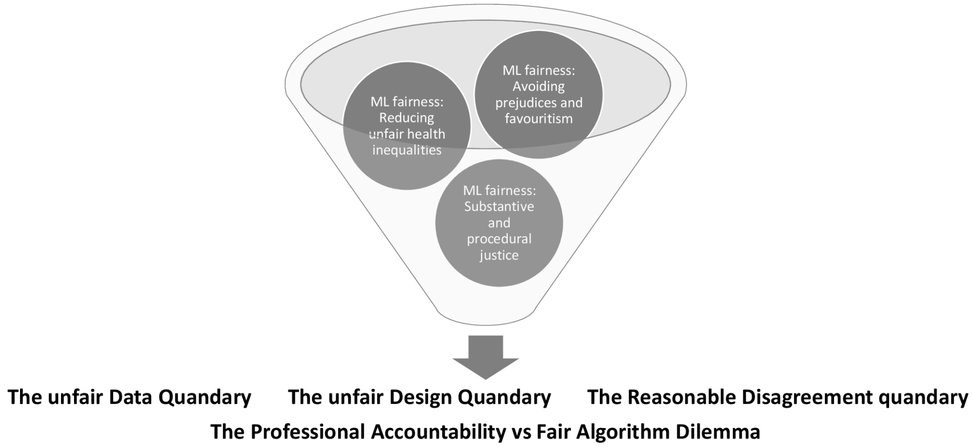

ML fairness: Reducing unfair health inequalities
ML fairness: Avoiding prejudices and favouritism
ML fairness: Substantive and procedural justice
The unfair Data Quandary
The unfair Design Quandary
The Reasonable Disagreement quandary
The Professional Accountability vs Fair Algorithm Dilemma

Figure 1 Three interrelated fairness quandaries and one fairness dilemma related to obtaining machine learning (ML) fairness in health are identified in this ethical analysis. A meaningful conceptualisation of ML fairness, which can be implemented to avoid inequitable patient outcomes, must reflect this complex, intertangled nexus of fairness concerns. This figure is made by the first author.

classifications or predictions.[17] Analogous to the animal brain, the mechanisms in deep learning are 'deep neural networks' consisting of hierarchically structured layers of 'neurons'. To work effectively, the neural networks are trained on vast data sets, which are sometimes labelled by humans (as in supervised learning) or they identify patterns in data sets on their own (as in unsupervised learning).[18] Due to its ability to identify pattern in vast data sets much faster and often more accurately than medical doctors, health professionals and scientists are able to, ML algorithms have the potential to make the detection, prediction and treatment of disease more effective.[17 19]

Substantial and procedural justice

Fairness can be analysed in terms of distinct principles of what justice requires (*substantive justice*) or in terms of the acceptability of how the decision is made (*procedural justice*).[20] The assumption behind procedural justice is that even though there may be widespread disagreement about what it would be just to do (eg, how to prioritise healthcare with resource scarcity), the affected parties may be expected to agree on what conditions must be in place to make the decision-making process fair.[21] Procedural justice requires, for example, that affected parties are treated equally by considering all interests at stake, and that decisions are based on reasons that individuals can recognise as relevant and reasonable.[21] Both versions of justice are relevant for fair decision making and how fairness issues come into play in relation to ML fairness.

Three quandaries and one dilemma

The unfair data quandary

The first quandary is related to biased data. This quandary states that groups of people not accurately represented in the training data of ML algorithms could receive diagnosis and treatment recommendations systematically imprecise in their disfavour. Since healthcare data typically emerge from contact with and/or use of the healthcare system, the extent to which people have *access* to healthcare will predict their inclusion in ML training data.

A conceptualisation of 'access to healthcare' can be divided into a supply side of the organised service and a demand side of patients' ability to benefit from the organised care. 'Access to healthcare' can be conceptualised across different phases involved in having a healthcare need met, that is, having a need, perceiving a need and desire for care, seeking healthcare, reaching healthcare services, using healthcare services and obtaining healthcare outcomes.[22] This broad approach to access to healthcare is useful for a nuanced investigation of where, when, how and by whom inequality in access can emerge under the impact of organised healthcare itself.

Healthcare services can uphold or reinforce social inequality in health if access to services requires capacities associated with socioeconomic conditions unequally distributed in the population. If the supply side is not carefully developed to meet the social and economic challenges related to people's abilities to reach and obtain care (eg, ability to pay or follow prescribed regimes, understanding of their own health or how the system works, cultural conflicts), data gathered from these services could be biased favouring those with the abilities to overcome barriers (eg, by paying for health insurance). Thus, unequal access skews the representativeness of big data gathered within the healthcare system to the advantage of those who have historically been able to use it. As this is the available data that ML algorithms are trained on, the detection of disease and clinical recommendations might not be equally apt for the groups that experience barriers in reaching, receiving and benefiting from care. This can be so if these latter groups overlap with relevant biological differences related to ethnical background, or if lifestyle issues related to socioeconomic challenges, impact the uptake of treatments. For the training data to be fair, the real-world conditions for access to healthcare must be equal in the sense that socioeconomic barriers do not prevent people from obtaining care.

The challenge to ensure fairness stemming from a lack of representative data is structural. Use of historically biased data combined with underdeveloped labelling creates racial biases in healthcare management of populations.[23 24] Space does not allow us to do justice to the vast literature on algorithm fairness and suggestions to mitigate algorithm biases. For a comprehensive overview, there is a framework proposed by Suresh and Guttag, which identifies the multiple sources of downstream harms caused by ML through data generation, model building, evaluation and data deployment, and also describes mitigation techniques for targeting the same sources.[2] As noted above, there are strong ethical and political calls to promote equal access to high-quality healthcare for all. To avoid a situation where ML unfairly maintains (or even reinforces) inequality in health outcomes, coordinated initiatives could be directed comprehensively towards identifying barriers and seeking innovative solutions to promote equal access to healthcare in the first place along all dimensions of supplying and demanding healthcare. Developers of ML systems, ethicists and funding bodies could join forces and gear attention towards mitigating the structural unfairness of unequal access to healthcare before addressing the inequitable outcome of this unfairness.

The unfair design quandary

Let us assume that comprehensive work has been done to ensure equal access to healthcare for all, which can enable fairness in algorithms deployed at the point of care. Now fairness is an issue about what kind of ML-based healthcare ought to be developed, that is, what kind of ML should be prioritised. How should this fairness aspect be ensured in the *design phase* when fair design then ideally must include broad oversight of consequences and justified priority-setting decisions *before* ML interventions have been developed and tested?

First, the ethical issues that arise from an ML system will depend on its practical application and purpose: is the system used for home monitoring, clinical decision support, improved efficiency and precision in testing, distribution and management of medicines, or something else? What kind of disease or ailment is being addressed? The ethical problems will be different and include different actors.

Next, one should ask: is ML needed, or do existing approaches work better? This is about the performance of ML, for example in terms of improved prediction.[25] But it is also about getting the process of interpreting what it means 'to work better', right. Who should decide that?

Depending on the problem being addressed, different actors will be involved. Design is not a linear process.[26] It depends on reiterated cycles of design, implementation, testing (including with other data), assessment and evaluation. This is even more so with ML algorithms, as they might display unpredictable outcomes (depending on input data, but also coding and algorithms). They therefore need constant human monitoring and assessment. The European Commission, for example, emphasise the need for stakeholder involvement throughout all the cycles.[27]

Potentially, these phases will involve inputs from people such as medical doctors, nurses, hospital administrators, health economists, other technical people and (ideally) the patients themselves. This then poses the question of the competencies that should enter into the design, implementation and testing phases, how they should be made to cooperate, and what kinds of expertise should count. Whose professional perspective may frame the initial understanding of the problem, what happens to dissenting voices, and what about patients' perspectives and autonomy? If these challenges to justice are not explicitly addressed, it might create an unfair design quandary . Procedural justice requires developing adequate and fair decision-making institutions for collaboration. This must be organised so all stakeholders can recognise them as being fair by including general requirements on transparency, reasonable justifications and opportunities for revision.[21] Still, figuring out how to best do so in these contexts requires more research.

The reasonable disagreement quandary

People are expected to disagree about principles of justice, what societal challenges one will trade off to improve people's health, and what opportunity costs one will accept to achieve health equality in the design process. How should such ethical disagreements be resolved? Procedural fairness addresses the moral equality of anyone involved in or being affected by a decision by arranging a decision-making process in a way that all can find acceptable. This means, for example, that all stakeholders must be included, allowing everyone to voice their concerns and listen to them, ensuring transparency of the rationales for the decision, and offering mechanisms to appeal.[21] In the case of designing and applying ML in medicine, there are multiple groups of experts, professions and other stakeholders that might play a central role in this kind of ethical deliberation, for instance medical doctors, nurses, hospital administrators, health economists, technicians, ML developers, patients, and the public in general. However, such inclusive deliberation might not be feasible to arrange every time an ML system is developed. The complex task of identifying, understanding and weighing all relevant medical, ethical, economical and societal issues to consider and reasonably justify what to prioritise in order to apply ML requires substantial technical and disciplinary expertise. Also, to ensure the prioritisation is adequately reflected in the design process, it is crucial to rely on ML experts and their interpretations when translating normative decisions into algorithms. This 'reasonable disagreement quandary' requires a fix in terms of fairness, but an overall fair decision-making process can be difficult to realise. Moreover, the fairness of leaving the decision to trained decision makers or technical experts and the substantial principles of justice they happen to hold, is also questionable.

More research is required to learn how to better maximise inclusiveness and transparency and monitor whether ethical and political prioritisations are captured in ML systems in a meaningful way. The aim should be to accommodate procedural fairness. However, realism is needed in identifying and articulating the limitations of such a fairness approach. A hybrid model of fairness based on substantial and procedural justice might emerge as a solution.

A final ethical dilemma

Let us, for the sake of argument, assume that the above quandaries are solved. Let us suppose that measures have been taken to ensure that the training data are not skewed, that adequate institutional conditions for collaboration between stakeholders and designers have been established, and that an acceptable model of procedural fairness has been developed. There is still, however, the following dilemma that needs to be addressed: while medical algorithms might improve fairness by eliminating biases that otherwise might affect the decisions of healthcare professionals and therefore result in more equitable access to healthcare services, they might also reduce the accountability of healthcare professionals for these same decisions. Algorithmic decision systems are built so it makes it difficult to determine why they do what they do or how they work. For example, neural networks that implement deep learning algorithms are large arrays of simple units, densely interconnected by very many links. During training, the networks adjust the weights of these links to improve performance, essentially deriving their own method of decision making when trained on a decision task. They therefore run independently of human control and do not necessarily provide an interpretable representation of what

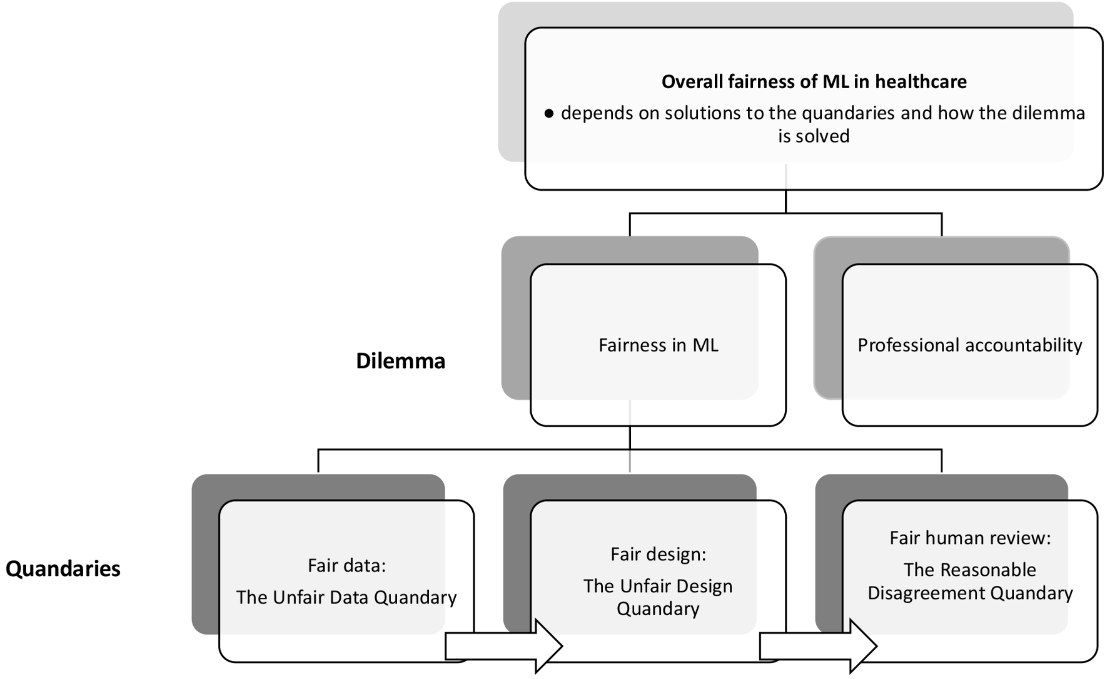

Overall fairness of ML in healthcare
● depends on solutions to the quandaries and how the dilemma is solved
Dilemma
Fairness in ML
Professional accountability
Quandaries
Fair data:
The Unfair Data Quandary
Fair design:
The Unfair Design Quandary
Fair human review:
The Reasonable Disagreement Quandary

Figure 2 Overall fairness of machine learning (ML) in healthcare depends on solutions to the three interrelated quandaries and the dilemma. This figure is made by the first author.

they do.[28 29] The problem with this is that *professional accountability* cannot be enforced without explainability. Professional accountability implies that it is justifiable to ask a healthcare professional to explain their actions and to clearly articulate and justify the decisions they have made. Providing such an explanation engenders trust in the process that led to the decision and confidence that the healthcare professional in charge of the process acted fairly and reasonably. It is true, as some have pointed out, that lack of explainability in medicine is not uncommon—sometimes it may be close to impossible to reconstruct the exact reasoning underlying the clinical judgement of a medical expert and there may be little knowledge of the causal mechanisms through which interventions work.[30] However, explainability is still important in some contexts, particularly in those requiring informed consent. In contexts where explainability *is* important, the potential opacity of ML algorithms suggests that some trade-off must be made between deferring to said algorithms (which might improve fairness but reduce accountability) and relying on human professional discretion (which might preserve accountability but increase the risk of biases). The dilemma is that neither option comes without ethical costs: either reduced accountability or (potentially) reduced fairness. Figure 2 shows how the quandaries and the dilemma are interrelated and part of a broad conceptualisation of ML fairness in healthcare.

CONCLUSION

We have demonstrated that operationalising fairness in ML algorithms in healthcare raises a whole host of fairness challenges across data, design and implementation biases, which all need to be solved before concluding that the algorithms are fair. Even if we have the ability to meet these challenges, we nevertheless face the problem of trading fair algorithms off against professional accountability. To avoid a rhetorical and insufficiently justified conception of fairness in ML technology, these fundamental and intangible challenges of fairness must be openly acknowledged and addressed. In addition, much more research on fair processes is called for to find ethically and politically sustainable responses to what fairness requires of ML algorithms employed in clinical care.

Acknowledgements The authors thank two anonymous reviewers for their valuable suggestions.

Contributors KB had the idea, drafted the first version and is acting as the guarantor. TG, EH and KR made substantive inputs to the draft and contributed to the revisions of the final manuscript.

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 No data are available.

ORCID iD

Kristine Bærøe http://orcid.org/0000-0002-4626-7232

REFERENCES

1 Reagan M. Understanding bias and fairness in AI systems, 2021. Available: https://towardsdatascience.com/understanding-bias-and-fairness-in-ai-systems-6f7fbfe267f3 [Accessed 03 Jul 2021].

2 Suresh H, Guttag JV. A framework for understanding unintended consequences of machine learning. *arXiv* 2019;2:190110002.

3 Mehrabi N, Morstatter F, Saxena N. A survey on bias and fairness in machine learning. *ArXiv* 2019:abs/1908.09635.

4 World Medical Association. Declaration of Geneva. Available: https://www.wma.net/policies-post/wma-declaration-of-geneva/2018 [Accessed 15 Jun 2020].

5 Daniels N. *Just health: meeting health needs fairly*. Cambridge University Press, 2007.

6 Sen A. Why health equity? *Health Econ* 2002;11:659–66.

7 Peter F. Health equity and social justice. *J Appl Philos* 2001;18:159–70.

8 Braveman PA, Kumanyika S, Fielding J, *et al*. Health disparities and health equity: the issue is justice. *Am J Public Health* 2011;101 Suppl 1:S149–55.

9 Marmot M, Friel S, Bell R, *et al*. Closing the gap in a generation: health equity through action on the social determinants of health. *Lancet* 2008;372:1661–9.

10 United Nations. *Transforming our world: the 2030 agenda for sustainable development*. New York: United Nations, Department of Economic and Social Affairs, 2015.

11 Bærøe K, Bringedal B. Just health: on the conditions for acceptable and unacceptable priority settings with respect to patients' socioeconomic status. *J Med Ethics* 2011;37:526–9.

12 Bringedal B, Bærøe K, Feiring E. Social Disparities in Health and the Physician's Role: A Call for Clarifying the Professional Ethical Code. *World Medical Journal* 2011;5:196–8.

13 Dahlgren G, Whitehead M. Policies and strategies to promote social equity in health. Background document to WHO - Strategy paper for Europe. *Arbetsrapport* 1991.

14 Wester G, Bærøe K, Norheim OF. Towards theoretically robust evidence on health equity: a systematic approach to contextualising equity-relevant randomised controlled trials. *J Med Ethics* 2019;45:54–9.

15 Rajkomar A, Hardt M, Howell MD, *et al*. Ensuring fairness in machine learning to advance health equity. *Ann Intern Med* 2018;169:866–72.

16 Wawira Gichoya J, McCoy LG, Celi LA, *et al*. Equity in essence: a call for operationalising fairness in machine learning for healthcare. *BMJ Health Care Inform* 2021;28:e100289.

17 Topol EJ. High-Performance medicine: the convergence of human and artificial intelligence. *Nat Med* 2019;25:44–56.

18 Franklin S. History, motivations, and core themes. In: *The Cambridge Handbook of artificial intelligence*, 2014: 15–33.

19 Rajkomar A, Dean J, Kohane I. Machine learning in medicine. *N Engl J Med Overseas Ed* 2019;380:1347–58.

20 Miller D. Justice. In: Zalta EN, ed. *The Stanford encyclopedia of philosophy*. 2017 Edition, 2021. https://plato.stanford.edu/archives/fall2017/entries/justice/

21 Daniels N, Sabin J. *Setting limits fairly: can we learn to share medical resources?* Oxford University Press, 2002.

22 Levesque J-F, Harris MF, Russell G. Patient-centred access to health care: conceptualising access at the interface of health systems and populations. *Int J Equity Health* 2013;12:18.

23 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.

24 Benjamin R. Assessing risk, automating racism. *Science* 2019;366:421–2.

25 Desai RJ, Wang SV, Vaduganathan M, *et al*. Comparison of machine learning methods with traditional models for use of administrative claims with electronic medical records to predict heart failure outcomes. *JAMA Netw Open* 2020;3:e1918962.

26 Stewart J, Williams R. 10. The wrong trousers? Beyond the design fallacy: social learning and the user. In: Rohracher H, ed. *User involvement in innovation processes strategies and limitations from a socio--technical perspective*. Munich: Profil Verlag, 2005.

27 Independent High-Level Expert Group on Artificial Intelligence (AI IHLEG). *Ethics guidelines for trustworthy AI*. Brussels: European Commission, 2019.

28 Burrell J. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data Soc* 2016;3:205395171562251–12.

29 de Fine Licht K, de Fine Licht J. Artificial intelligence, transparency, and public decision-making. *AI Soc* 2020;35:917–26.

30 London AJ, Intelligence A. Artificial intelligence and black-box medical decisions: accuracy versus Explainability. *Hastings Cent Rep* 2019;49:15–21.

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 health-care 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.

BMJ Health & Care Informatics

Global disparity bias in ophthalmology artificial intelligence applications

Luis Filipe Nakayama ,[1] Ashley Kras,[2] Lucas Zago Ribeiro,[1] Fernando Korn Malerbi ,[1] Luis Salles Mendonça,[1,3] Leo Anthony Celi ,[4,5] Caio Vinicius Saito Regatieri,[1] Nadia K Waheed[3]

To cite: Nakayama LF, Kras A, Ribeiro LZ, *et al.* Global disparity bias in ophthalmology artificial intelligence applications. *BMJ Health Care Inform* 2022;**29**:e100470. doi:10.1136/bmjhci-2021-100470

Received 11 August 2021
Accepted 03 March 2022

[1]São Paulo Federal University, São Paulo, SP, Brazil
[2]Retinal Imaging Lab, Harvard University, Cambridge, Massachusetts, USA
[3]Tufts Medical Center, New England Eye Center, Boston, Massachusetts, USA
[4]Medicine, Beth Israel Deaconess Medical Center, Boston, Massachusetts, USA
[5]Laboratory for Computational Physiology, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA

Correspondence to
Luis Filipe Nakayama;
nakayama.luis@gmail.com

Machine learning (ML) is a branch of artificial intelligence (AI) that performs a classification, prediction and/or optimisation task. Similar to brain neurons, neural networks output a label after multiple information layers connection, resembling human thinking.[1]

AI is already influencing care in many areas, such as radiology, pathology, dermatology and ophthalmology. In ophthalmology, a variety of multimodal imaging examinations are fundamental in the screening, diagnosis and monitoring of diseases and provide data input for AI development.[2] Some applications, such as the IDx Technologies (Coralville, USA) which was approved by the Food and Drug Administration 3 years ago, are already used in clinical practice as a screening tool.[2 3] Surprisingly, algorithms can even predict gender, age and cardiovascular risk through retinal images.[2 4 5] AI may reduce subjectivity and interobserver disagreement in clinical practice.[1]

Especially in low-income (LIC) and low-to-medium-income countries (LMIC), preventable blindness causes such as diabetic retinopathy (DR) and age-related macular degeneration could be prevented with screening programmes, home monitoring systems or telemedicine. AI-based screening could systematise screening and improve eye care in remote areas.[6]

ML requires high-quality, well-labelled, representative and large datasets, but at present, ophthalmological ML-ready datasets are only available in a few countries. One hundred seventy-two countries do not have representation in training and validation cohorts.[7]

Although data from all world countries are a distant goal, equivalent representation of all continents, ethnicities and the maximum number of countries is desired to reduce ML bias. Demographic information and other social determinants of health are typically not contained in these datasets, making it challenging to interrogate algorithms for bias.[7 8] High-quality data are also fundamental for environmental-specific algorithm validation, which is essential before AI implementation.

Available automated DR algorithm performance varies considerably in performance in the real world due to limited training data, including heterogeneity in disease presentations and suboptimal image quality.[9] In addition, diverse sociodemographic and ethnic representation are necessary if generalisability is a goal.[8]

In LICs and LMICs, there is a growing gap between the ophthalmologist workforce and the population size. Two-thirds of ophthalmologists live in only 17 countries and in those countries, most practice in the urban centres.[10] AI applications can expand access to eye care and may reduce preventable blindness, which is currently 80% of cases.

In addition to diversifying datasets to build AI technology in healthcare, we must invest in building capacity for health informatics and data science across countries. International collaboration between research groups should be incentivised to narrow disparities in AI research in order to reduce world blindness.

Twitter Leo Anthony Celi @MITCriticalData

Contributors All authors designed and cowrote the draft of the paper and critically reviewed and edited the final draft. LFN and AK took responsibility for article concept.

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval UNIFESP Ethics Institutional Review Board number: CAAE 33842220.7.0000.5505 / n:0698/2020.

Provenance and peer review Not commissioned; externally peer reviewed.

ORCID iDs

Luis Filipe Nakayama http://orcid.org/0000-0002-6847-6748
Fernando Korn Malerbi http://orcid.org/0000-0002-6523-5172
Leo Anthony Celi http://orcid.org/0000-0001-6712-6626

REFERENCES

1 He M, Li Z, Liu C, *et al*. Deployment of artificial intelligence in real-world practice: opportunity and challenge. *Asia Pac J Ophthalmol* 2020;9:299–307.
2 Kras A, Celi LA, Miller JB. Accelerating ophthalmic artificial intelligence research: the role of an open access data Repository. *Curr Opin Ophthalmol* 2020;31:337–50.
3 Md A, PT L, M B, *et al*. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *Yearbook of Paediatric Endocrinology* 2019.
4 Poplin R, Varadarajan AV, Blumer K, *et al*. Prediction of cardiovascular risk factors from retinal fundus Photographs via deep learning. *Nat Biomed Eng* 2018;2:158–64.
5 Korot E, Pontikos N, Liu X, *et al*. Predicting sex from retinal fundus Photographs using automated deep learning. *Sci Rep* 2021;11:10286.
6 Xie Y, Gunasekeran DV, Balaskas K, *et al*. Health economic and safety considerations for artificial intelligence applications in diabetic retinopathy screening. *Transl Vis Sci Technol* 2020;9:22.
7 Khan SM, Liu X, Nath S, *et al*. A global review of publicly available datasets for Ophthalmological imaging: barriers to access, usability, and generalisability. *Lancet Digit Health* 2021;3:e51-e66.
8 Mitchell WG, Dee EC, Celi LA. Generalisability through local validation: overcoming barriers due to data disparity in healthcare. *BMC Ophthalmol* 2021;21:228.
9 Lee AY, Yanagihara RT, Lee CS, *et al*. Multicenter, head-to-head, real-world validation study of seven automated artificial intelligence diabetic retinopathy screening systems. *Diabetes Care* 2021;44:1168–75.
10 Resnikoff S, Lansingh VC, Washburn L, *et al*. Estimated number of ophthalmologists worldwide (international Council of ophthalmology update): will we meet the needs? *Br J Ophthalmol* 2020;104:588–92.